Towards Safe and Trustworthy Embodied AI: Foundations, Status, and Prospects

Xin Tan[1,2*], Bangwei Liu[1,2*], Yicheng Bao[1,2], Qijian Tian[1], Zhenkun Gao[1,2], Xiongbin Wu[1], Zhihao Luo[1,2], Sen Wang[1,2], Yuqi Zhang[1], Xuhong Wang[1§], Chaochao Lu[1§†], Bowen Zhou[1,3§‡]

[1]Shanghai Artificial Intelligence Laboratory
[2]East China Normal University
[3]Tsinghua University

GitHub Project Page

2025-09-12

Abstract

The increasing autonomy and physical capability of Embodied Artificial Intelligence (EAI) introduce critical challenges to safety and trustworthiness. Unlike purely digital AI, failures in perception, planning, or interaction can lead to direct physical harm, property damage, or the violation of human safety and social norms. However, current EAI foundation models disregard the risks of misalignment between the model capabilities and the safety and trustworthiness competencies. Some works attempt to address these issues, however, they lack a unified framework capable of balancing the developmental trajectories between safety and capability. In this paper, we first comprehensively define a new term *safe and trustworthy EAI* by establishing an L1-L5 levels framework and proposing ten core principles of trustworthiness and safety. To unify fragmented research efforts, we propose a novel, agent-centric framework that analyzes risks across the four operational stages of an EAI system. We systematically review state-of-the-art but fragmented solutions, benchmarks, and evaluation metrics, identifying key gaps and challenges. Finally, we identify the need for a paradigm shift away from optimizing isolated components towards a holistic, cybernetic approach. We argue that future progress hinges on engineering the closed-loop system of the agent (Self), its environment (World), and their dynamic coupling (Interaction), paving the way for the next generation of truly safe and trustworthy EAI.

[*]These authors contributed equally. [§]Corresponding Author. [†]Project Leader. [‡]Scientific Leader.

Contents

1 Introduction

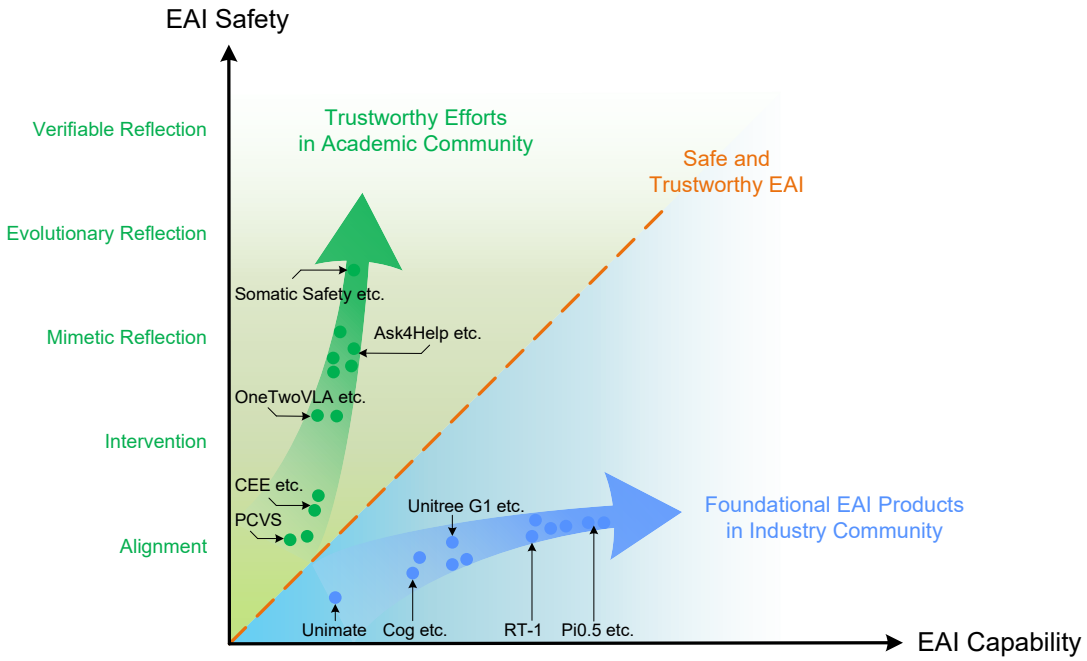

Figure 1 The divergence between capability and safety in the Embodied AI landscape. The horizontal axis charts general EAI capability. The vertical axis represents our proposed EAI Safety maturity model, building upon the general AI Safety levels introduced in R²AI [184]. We observe two divergent trends: current industry products (blue) are advancing rapidly in capability but lag in safety maturity. Conversely, academic research (green) is exploring higher safety levels but is often concentrated on less capable systems. Our work aims to chart a course toward the ideal trajectory of a Safe and Trustworthy EAI (orange), bridging the critical gap between these two trends.

Embodied Artificial Intelligence (EAI), the integration of perception, cognition, and physical interaction within a closed-loop agent—has seen rapid progress in recent years, driven by advances in robotics, computer vision, natural language processing, and large multi-modal language model [36]. Embodied agents are increasingly deployed in real-world scenarios, including household service robots [89], autonomous vehicles [240], assistive healthcare devices [107], and interactive learning environments.

As these agents become more complex and autonomous, they also bring serious safety and trust challenges. Unlike traditional AI systems that work in fixed or virtual settings, embodied agents must deal with changing, uncertain environments and make physical contact with people and objects [53]. A single mistake in perception, planning, or following instructions can lead to harmful behavior, ethical problems, or security risks—such as hurting a person or being tricked by an attacker [61]. Such failures raise critical concerns about the safety and trustworthiness of embodied AI. However, current EAI foundation models are undergoing rapid advancement, yet they entirely disregard the risks of misalignment between the model capabilities and the safety and trustworthiness competencies.

As shown in Figure 1, representative works such as Helix [46], $\pi 0.5$ [72], and RT-2 [254] have established themselves as foundational EAI models. However, none of them explicitly declare the implementation of powerful safety alignment mechanisms, which constitutes an irresponsible approach for physical-level AI systems. Conversely, while the academic community has produced fragmented research

efforts addressing EAI safety and trustworthiness concerns, these investigations consistently fail to integrate safety capabilities into widely used foundational models. We contend that the fundamental issue stems from the absence of a coherent developmental framework for safety and trustworthy EAI research that can be harmoniously aligned with the advancement of EAI capability.

The academic community has indeed produced a growing body of work aimed at addressing the safety and trust challenges of embodied intelligence. However, as illustrated by the scattered distribution of research efforts in Figure 1, these investigations are largely fragmented and focus on narrow, isolated aspects of the problem. Much of the research is concentrated at the lower levels of our proposed safety maturity model (as detailed in Chapter 2). For instance, a significant cluster of work explores Alignment, proposing new benchmarks to test for harmful instruction following [101] or developing new defense mechanisms against prompt-based attacks [228]. Another active area is Intervention, with studies focusing on enabling agents to ask for help when uncertain [175] or designing novel architectures for internal safety verification [80]. A third cluster is emerging around Mimetic Reflection, investigating how imitation learning can be made safer or how to ensure somatic safety during physical interaction [15]. While these contributions are individually valuable, they rarely connect to form a cohesive, systemic approach. This fragmentation highlights a critical gap: the field lacks a unified developmental framework that would allow these disparate research threads to be integrated, preventing the community from building embodied systems that are holistically safe and trustworthy.

Ensuring that embodied agents operate safely is crucial for their real-world application. However, as our review shows, current research on this topic is often fragmented. Many studies focus on solving specific problems, such as making a robot's grip more reliable or defending it from a certain type of attack. Researchers often address individual principles like reliability or security in isolation, without a clear understanding of how these different aspects connect and interact within a complete embodied system. Therefore, the research community and industry urgently need a shared framework that defines what "safe and trustworthy" means for an embodied agent. Such a framework is essential for systematically measuring progress, identifying critical research gaps, and guiding the development of holistically safe systems.

To address this need for a unified structure, this paper makes three primary contributions.

- First, we formally introduce and define the concept of **Safe and Trustworthy Embodied AI**, establishing it as a holistic field of study that integrates both the internal reliability of an agent and its external safety in the physical world.
- Second, we propose a five-level maturity model for "Make Safe EAI". This model provides the first clear roadmap for the field's evolution, charting a course from reactive, externalized safety measures to proactive, intrinsically safe and resilient systems.
- Third, we present a comprehensive framework of ten core principles, organized under the two primary dimensions of Trustworthiness and Safety. This framework serves as a systematic tool for analyzing risks, classifying existing research, and identifying critical gaps, guiding the development of agents that are not just capable, but fundamentally safe and dependable.

How our focus differs from prior surveys? This paper is not only a survey, since we take more efforts to define a new concept (i.e., Safe and Trustworthy Embodied AI) with the five-level maturity

model. Existing surveys on trustworthy AI, such as those by Li et al.[90], offer valuable high-level principles, but their generalist approach does not fully capture the unique challenges of embodiment. Embodied intelligence is not merely an application of AI; it is a complex system where software, hardware, and physical interaction are deeply intertwined. Similarly, surveys on LLM safety[66, 69] focus on risks in the text and image domains, overlooking the critical new failure modes that emerge when an LLM's outputs are translated into physical actions. Even within the embodied AI literature, prior surveys have tended to focus on specific, isolated aspects of trustworthiness, such as manipulation reliability [248, 210], explainability [196], or ethics [17]. The most closely related work, a survey by Neupane et al. [134], provides an excellent taxonomy of security threats based on attack surfaces, ethics and human-robot interaction. However, it overlooks the profound paradigm shift where Large Language Models have become the "brain" for general-purpose embodied agents.

Structure of this Paper This paper is organized as follows. We begin in Chapter 2 by introducing our proposed safe and trustworthy maturity model, the "Five Levels of Make Safe EAI". This chapter provides a high-level roadmap for the field's evolution, charting a course from reactive, externalized safety measures to proactive, intrinsically safe and resilient systems. Chapter 3 establishes our foundational framework. It begins by distinguishing between the two core dimensions of Safety and Trustworthiness, and then details the ten core principles that guide our analysis throughout the paper. With this framework in place, Chapter 4 provides a comprehensive review of the current research landscape, organized along the agent's operational workflow: Instruction, Perception, Planning, and Interaction. Within each stage, we analyze existing work through the lens of the safety and trustworthiness principles defined earlier. Building upon this analysis, the subsequent chapters present our primary contributions and forward-looking positions. Chapter 5 and 6 then delves into the critical role of benchmarks and simulators, evaluating them not just on their functionality, but on their capacity to support trustworthiness research. Finally, Chapter 7 presents our position on the future of the field, proposing a closed-loop, cybernetic agenda for research centered on three pillars: high-fidelity environments (the World), self-evolving agents (the Self), and seamless coordination (the Interaction). We conclude by summarizing our findings and reiterating our call for a more holistic, systems-level approach to building the next generation of trustworthy embodied AI.

2 Levels of Safe and Trustworthy Embodied AI

2.1 From "Making Embodied AI Safe" to "Making Safe Embodied AI"

The discourse surrounding AI safety is undergoing a fundamental shift, moving from a reactive posture to a proactive one. This evolution is best captured by the distinction between two philosophies: "making an embodied AI safe" versus "making a safe embodied AI." The former represents a traditional, post-hoc approach where safety mechanisms are treated as external add-ons or "guardrails" bolted onto a pre-existing, powerful but untrusted intelligence. This approach, while valuable, is inherently limited, as it contains risk without fundamentally changing the agent's nature.

In contrast, "making a safe embodied AI" is a proactive, safe-by-design paradigm. It posits that safety cannot be an afterthought; it must be a core competency embedded within the agent's architecture, reasoning, and learning processes from the very beginning. Safety is not a constraint on capability but is, in fact, a capability itself. While several frameworks exist for grading robot capabilities, such as the

five-level autonomy standard (IR-L0 to IR-L4) detailed in recent surveys [108], these primarily focus on task performance, autonomy, and intelligence. They answer "what can the robot do?", but not "how is the robot built to be safe?".

To fill this gap, we propose a new maturity model specifically for "Make Safe Embodied AI", organized into five levels (L1-L5), which is build upon the foundational "Make Safe AI" framework proposed in R²AI [184]. As shown in Figure 2, this framework includes two complementary pillars: Resistance, the ability to withstand and prevent known threats, and Resilience, the capacity to recover, adapt, and strengthen when facing novel or unforeseen challenges. Resistance anchors the foundational layers (L1-L2), while Resilience drives the advanced layers of adaptation and self-improvement (L3-L5), providing a clear roadmap toward truly robust and provably safe systems.

2.2 The Five Levels of "Make Safe EAI"

Our five-level model adapts the core principles of the R²AI framework [184] to the embodied domain. The following sections will now detail each of these levels shown in Figure 2, illustrating the progressive journey from foundational, data-driven alignment to adaptive, and ultimately, verifiable systems. Each level represents a significant leap in an agent's intrinsic safety capabilities, defining a clear and measurable pathway for future research and development.

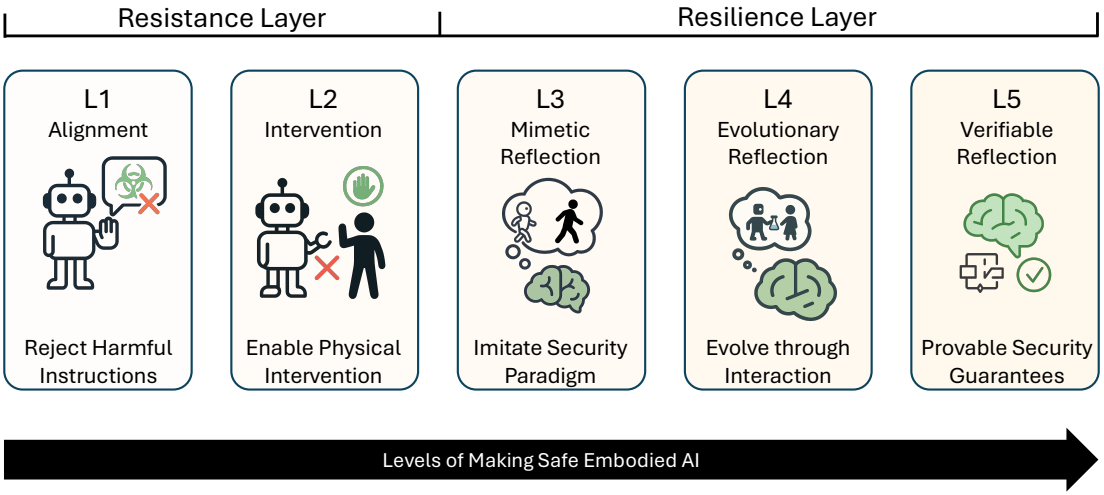

Figure 2 The five levels of our "Make Safe EAI" maturity model for embodied intelligence. This framework builds upon the core concepts of the general AI safety levels introduced in R²AI [184], adapting them to the unique challenges of the physical world. The levels are organized into two foundational pillars: **Resistance** (L1-L2), which focuses on withstanding known threats, and **Resilience** (L3-L5), which enables adaptation and recovery from novel challenges.

2.2.1 L1: Alignment - Foundational Resistance

At this foundational level, safety is achieved by aligning the agent's behavior with basic human values and safety norms through large-scale, data-driven training. This forms the first layer of Resistance. An L1 agent is not intrinsically reasoning about safety; rather, it has learned to associate certain patterns in instructions, perceptions, and planned actions with "safe" or "unsafe" outcomes based on its training

data. For an embodied agent, this means being able to refuse clearly harmful instructions, such as those cataloged in benchmarks like AGENTSAFE, with a high refusal rate [101]. The key technologies are Instruction-Tuning and Reinforcement Learning from Human Feedback (RLHF) [27, 140, 13]. However, this alignment is often "skin-deep" and brittle, vulnerable to sophisticated jailbreaking attacks that exploit the gap between its learned correlations and a true understanding of the physical world [243].

2.2.2 L2: Intervention - Resistance through Oversight

Level 2 enhances Resistance by enabling robust external oversight and interruption. The core principle is that even an aligned agent can make mistakes, and thus, a human must always be in a position to intervene. For an embodied agent, this extends beyond a simple emergency stop button. It requires the agent's decision-making process to be transparent enough for a human to understand its intent before a potentially harmful physical action occurs. Key technologies include Explainable AI (XAI) to articulate the rationale behind a plan [196], and systems that explicitly visualize an agent's future trajectory or intentions [34]. An L2 agent is designed to respond reliably to interruption commands [139], making it a trustworthy collaborator under human supervision. Its limitation is its reliance on constant human vigilance, which does not scale to fully autonomous operation.

2.2.3 L3: Mimetic Reflection - Foundational Resilience

This level marks the transition from purely resisting threats to building Resilience. The core idea is that an agent can learn to be safer by reflecting upon and internalizing proven safe behaviors. Instead of just being programmed with what not to do, an L3 agent learns how to perform tasks safely by imitating validated safe behavior templates, whether from human demonstrations or a curated library of best practices. For an embodied agent, this means learning complex manipulation or navigation skills by observing experts, a process often facilitated by imitation learning [167, 168] and behavioral cloning [148]. Through mimetic reflection, safe conduct becomes an internalized part of the agent's expertise, allowing it to generalize safe behaviors to similar but unseen situations. However, its resilience is limited by the diversity of its demonstrated knowledge; it struggles to handle truly novel scenarios that have no precedent in its experience [162].

2.2.4 L4: Evolutionary Reflection - Adaptive Resilience

At Level 4, Resilience becomes an adaptive, autonomous process. An L4 agent develops self-improvement mechanisms, allowing it to learn and refine its safety strategies through continuous interaction with the physical world. This is where the agent's entire perception-planning-interaction loop becomes a lifelong learning system. It learns from the rich feedback of its own physical experiences—a near-miss, an unexpected sensor reading, or an action that did not produce the expected physical result. Key technologies include continual learning to adapt to new threats without catastrophic forgetting [142], and the ability to perform self-generated "red teaming" to proactively discover and patch its own vulnerabilities [145]. This evolutionary capability allows the agent to build resilience against novel and long-term risks, but its learning is still empirical and cannot provide prior guarantees of safety.

2.2.5 L5: Verifiable Reflection - Guaranteed Resilience

Level 5 represents the pinnacle of "Make Safe AI", where resilience is fortified by provable guarantees grounded in control theory. An L5 agent can not only adapt but also reflect on its own dynamic stability, providing formal assurances about its behavior. For an embodied agent, this means its actions are governed by a control law that is verifiably safe. This requires a deep integration of learning with control-theoretic methods, such as reachability analysis, which can compute all possible future states to verify safety [4, 188]. Neuro-symbolic architectures are a key enabling technology, allowing for the formal verification of learned policies against symbolic safety constraints [52]. While this level offers the highest form of trustworthiness, it acknowledges the profound sim-to-real gap [86]: the proof is only as valid as the underlying dynamic model of the world, making the fidelity of that model a critical research frontier.

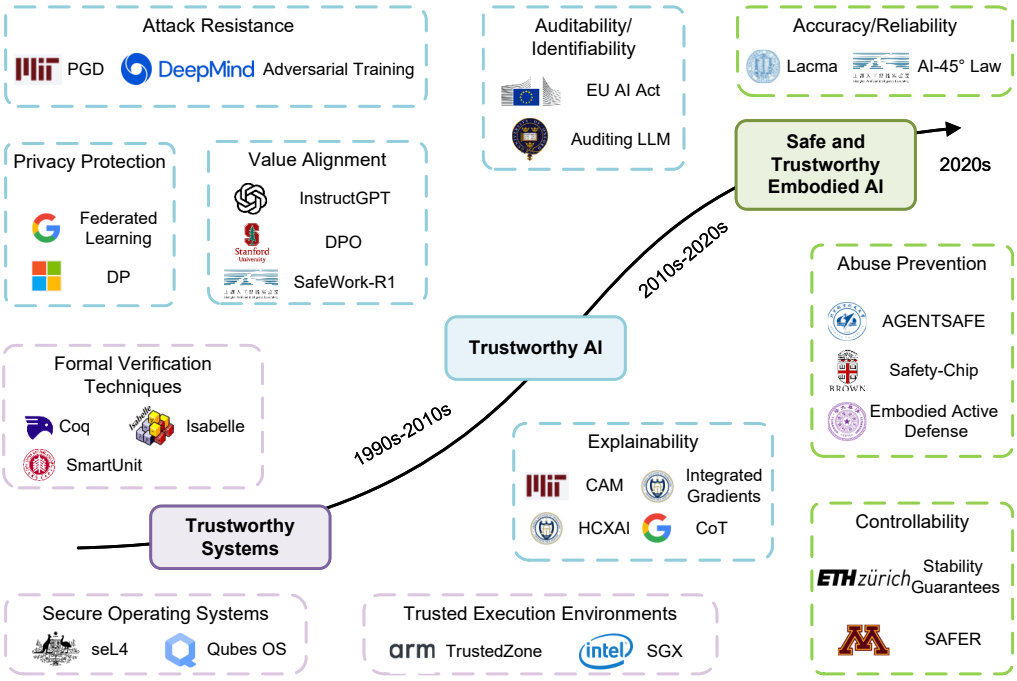

Figure 3 A timeline illustrating the evolution of trustworthy computing. The figure charts the progression from foundational systems to the cutting-edge frontier of embodied AI. Different research domains are distinguished by color: foundational concepts of **Trustworthy Systems** are highlighted in purple, core principles of **Trustworthy AI** are in blue, and the specific challenges of **Safe and Trustworthy Embodied AI** are marked in green.

3 Principles of Safe and Trustworthy Embodied AI

The principles governing safe and trustworthy systems have evolved in lockstep with the advancement of computing itself. As illustrated in Figure 3, this journey began with a focus on Trustworthy Systems, where foundational research in areas like Secure Operating Systems and Formal Verification established the bedrock of reliable computing. With the rise of machine learning, the focus shifted to Trustworthy

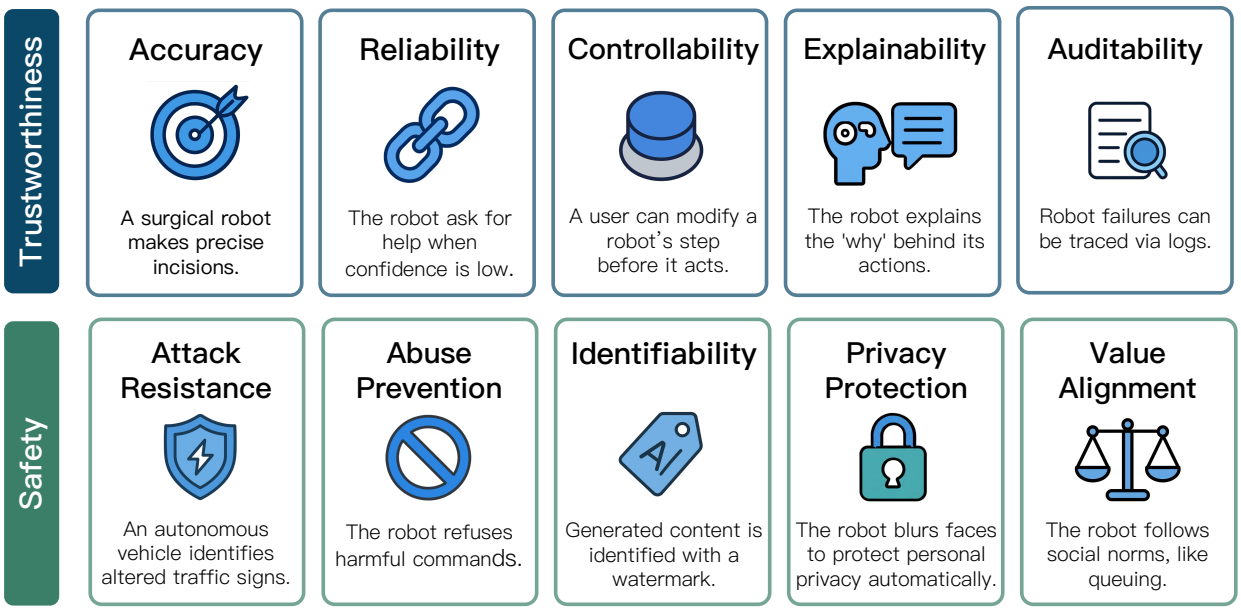

Figure 4 An overview of the ten core principles for trustworthy and safe embodied AI. The framework is divided into Trustworthiness (top row), which governs the agent's competence and predictability, and Safety (bottom row), which governs its capacity for harm avoidance and adherence to norms. Each card provides a concrete example to illustrate the principle in practice.

AI, introducing new principles such as Attack Resistance against adversarial examples, Explainability for opaque models, and Value Alignment for large language models.

Today, as AI moves from the digital area into the physical world, we are entering a new era: **Safe and Trustworthy Embodied AI**. This paradigm shift introduces unprecedented challenges that require an expanded set of principles. The direct physical interaction of embodied agents necessitates a renewed and urgent focus on principles like Abuse Prevention and Controllability, as failures are no longer confined to virtual errors but can result in tangible harm.

To address this new frontier, we must establish a clear set of guiding principles. This chapter introduces a comprehensive framework that distinguishes between two indispensable dimensions: Safety and Trustworthiness. We define these dimensions and then break them down into ten core principles: five for trustworthiness and five for safety, that holistically address the challenges of this new era. This framework provides a structured way to analyze the challenges and evaluate the progress in creating responsible embodied AI. Figure 4 provides an overview of these ten principles.

3.1 Safety and Trustworthiness: Two Indispensable Dimensions

While often used interchangeably, safety and trustworthiness represent two distinct but complementary aspects of a responsible embodied agent.

Safety focuses on *harm avoidance*. It is the absolute baseline requirement for any system that interacts with the physical world. The core question for safety is: "Will the agent, directly or indirectly, cause unacceptable physical, psychological, or property damage?" This dimension is primarily concerned with preventing negative outcomes and protecting against external threats.

Trustworthiness focuses on *confidence building*. It is the user's belief that an agent will consistently demonstrate competence, reliability, and integrity. The core question for trustworthiness is: "Can I rely on this agent to perform its tasks correctly and predictably, and if it fails, will I understand why?" This dimension is about ensuring positive and predictable performance, which builds the foundation for human-agent collaboration.

3.2 Five Principles of Trustworthiness

Trustworthiness is built upon the agent's internal capabilities and its ability to interact with the world in a predictable and understandable manner.

3.2.1 Accuracy

Accuracy is the principle that an agent's understanding and actions align precisely with reality and user intent. It has two components. First, **content accuracy** refers to the agent's ability to correctly perceive the world and understand the user's commands. For example, an agent must accurately identify objects in its environment and not "hallucinate" non-existent ones [79]. Second, **behavioral accuracy** means the agent's physical actions precisely match its intended task goals. If asked to place a cup on a coaster, its movements must be precise enough to succeed, demonstrating a true unity of knowledge and action [191, 174].

3.2.2 Reliability

Reliability is the agent's ability to maintain its performance consistently across different situations, especially in the face of uncertainty. A key aspect is **predictable performance**, where the agent delivers consistent results for repeated tasks. What's more, reliability includes robustness, which is the capacity to handle unexpected disturbances [147], such as sensor noise or slight changes in the environment. A reliable agent also understands its own capability boundaries. It knows what it can and cannot do, and it can perform a graceful degradation by safely stopping its task or asking for human help when it encounters a situation beyond its abilities [175].

3.2.3 Controllability

Controllability ensures that the agent's operations and associated risks remain under human oversight. This involves the ability to predict and assess risks before they occur, allowing for preventive measures. It also requires that the agent has the built-in capacity to avoid causing harm, even if its primary task logic fails [60]. Crucially, controllability demands mechanisms for effective human intervention, such as an emergency stop button or a clear process for a human to take over control, ensuring that a human operator is the ultimate authority in any situation [139].

3.2.4 Explainability

Explainability is the principle that an agent's decisions should be understandable to humans. This is essential for debugging, building trust, and enabling meaningful collaboration. It involves post-hoc explanation, where the agent can answer "Why did you do that?" in a clear, intelligible way after an action is complete [124]. For instance, an agent might explain it took a longer path to avoid a slippery

floor. Explainability also includes the real-time presentation of intent, where the agent visualizes its internal state or future plan, allowing users to understand its reasoning as it happens [196, 34].

3.2.5 Auditability

Auditability ensures that an agent's actions can be reviewed and analyzed after the fact, which is critical for accountability and learning from failures. The first component is **traceability**, which is the ability to reconstruct the agent's decision-making process by tracing the flow of information and logic that led to a specific action [92]. The second component is the availability of **verifiable evidence**. This requires the system to maintain secure, unalterable logs of its sensor data, internal states, and actions, creating a "black box" recorder that can be used for incident investigation [129].

3.3 Five Principles of Safety

Safety principles focus on protecting the agent and its environment from harm, particularly from intentional threats and unethical behaviors.

3.3.1 Attack Resistance

Attack resistance is the principle of protecting the agent's entire operational pipeline from malicious interference. This includes defending against external attacks that target its perception, cognition, or actuation. For example, an agent must resist adversarial attacks on its perception, where manipulated sensor data (e.g., a sticker on a stop sign) could cause it to misinterpret the world [40]. It must also defend against jailbreak attacks on its instruction understanding, where crafted prompts trick the agent into performing harmful actions [243]. Finally, its internal reasoning and physical actuators must be protected from being hijacked or manipulated by an attacker [221].

3.3.2 Abuse Prevention

Beyond defending itself, a safe agent must also prevent itself from being used as a tool for harm. Abuse prevention is the principle of identifying and refusing to carry out commands that have a malicious intent. This requires the agent to have a high-level understanding of the potential consequences of its actions. For example, if commanded to "use this hammer to break the window," the agent should recognize the destructive nature of the request and refuse to comply, distinguishing it from a benign command like "use this hammer to hang a picture" [117]. Such refusal capabilities are central to alignment research[10, 146], and are increasingly tested in large models through safety evaluations and system cards [2].

3.3.3 Identifiability

Identifiability ensures that an agent's presence and actions are clearly distinguishable from those of a human. This principle helps to prevent deception and manage social expectations. **Content identifiability** means that any language or content generated by the agent should be clearly marked as AI-generated, as emphasized in regulatory frameworks such as the EU AI Act [39]. **Behavioral identifiability** means the agent's physical form or behavior should make it clear that it is a robot. For example, a service robot in a public space might have a distinct appearance or an indicator light

to signal its autonomous operation [113], preventing situations where people might mistake it for a human-operated device, which is also emphasized in standards like IEEE 7001 [11].

3.3.4 Privacy Protection

Embodied agents, with their mobile sensors, operate in sensitive environments like homes and hospitals. The principle of privacy protection requires that they handle personal data with the utmost care. This includes data minimization, where the agent only collects the sensory information necessary for its task, as emphasized in the European Union's General Data Protection Regulation (GDPR)[159] and the extitIEEE Standard for Data Privacy Process (IEEE 7002-2022) [138]. It also involves the secure storage, transmission, and access control of any collected data, ensuring that private information, such as maps of a home or recordings of conversations, is protected from unauthorized access [93].

3.3.5 Value Alignment

Value alignment is the principle that an agent's behavior must align with human ethics and social norms. The most critical component is fairness, ensuring the agent does not make decisions that discriminate against any group of people [119]. It also includes adherence to social norms, such as respecting personal space or not interrupting conversations. Finally, it involves the capacity for ethical decision-making, where the agent can navigate moral dilemmas and make choices that reflect widely held societal values, such as prioritizing human well-being above all else [5].

3.4 Current Research Trends

To provide a quantitative snapshot of the current research landscape, we conduct a literature review of recent papers in embodied AI safety and trustworthiness. We categorize each paper based on the primary principle it addresses and the specific AI capability it focuses on. The results, summarized in Figure 5, reveal the distribution of research efforts across our proposed framework and highlight both areas of intense focus and those that are currently underexplored.

4 Research in Safe and Trustworthy EAI

4.1 Workflow of Embodied AI

Embodied AI can be understood through the perspective of four key stages that define how an intelligent agent interacts with its environment: instruction understanding, environment perception, action planning, and physical interaction. This differs from traditional robotic workflows, which typically focus on the three fundamental primitives: sense, plan, and act, as described by Murphy et al. As large language models (LLMs) mature and are integrated into embodied intelligence systems, they enable these embodied agents to process various user instructions and intentions, completing tasks autonomously based on the input they received. With this advancement, the embodied agent will be able to accept various instructions from humans, plan its own actions through perception of the environment, and ultimately interact with the real-world environment and humans, which is a more complex system compared with robotics. Figure 7 illustrates the four-stage workflow of embodied

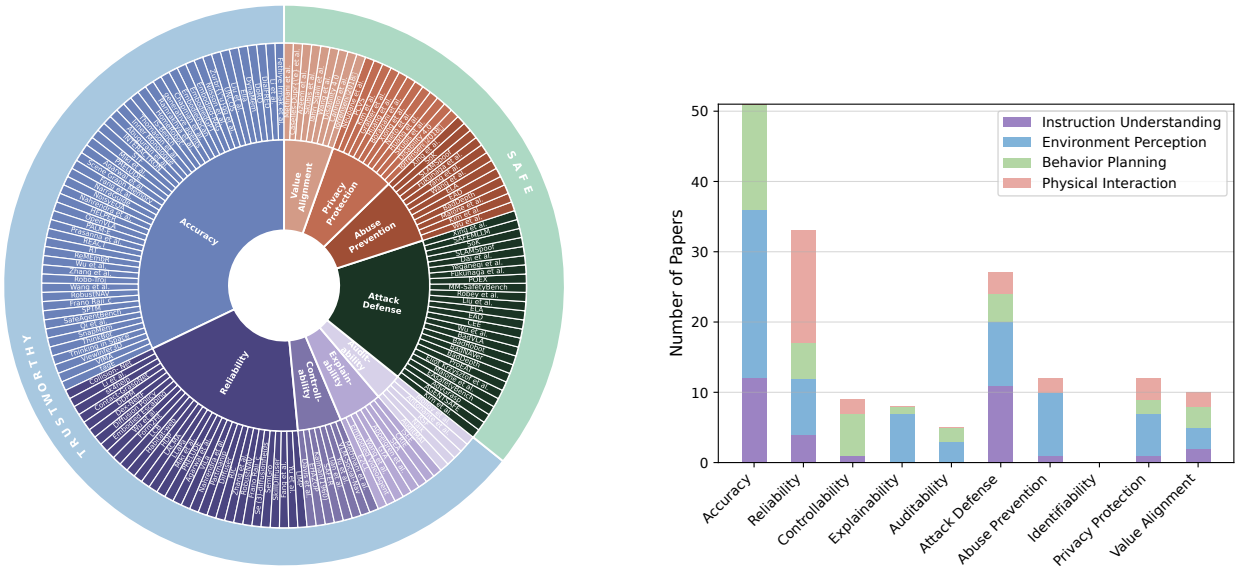

Figure 5 A quantitative overview of the research landscape for safe and trustworthy embodied AI. **Left:** A sunburst chart showing the hierarchical structure of the ten principles, grouped into Trustworthiness and Safety. **Right:** A stacked bar chart quantifying the number of reviewed papers for each principle, broken down by the AI capability they target (Instruction Understanding, Environment Perception, Behavior Planning, or Physical Interaction). The data indicates a strong research concentration in Accuracy, Reliability, and Attack Resistance, while principles such as Auditability and Identifiability are comparatively underexplored.

intelligence, highlighting the relationships between each phase and their contributions to a reliable and secure system.

To systematically organize the research in safe and trustworthy embodied AI, we have constructed a detailed literature taxonomy based on the four core stages proposed above: instruction understanding, environment perception, action planning, and physical interaction. Within each stage, this taxonomy further categorizes existing works according to key principles of safety and trustworthiness, such as value alignment, privacy protection, attack resistance, and reliability. The overall structure of this taxonomy, along with detailed citations to the relevant literature, is presented in Figure 6. The following subsections will adhere to this framework, delving into the specific challenges and research advancements within each stage.

Instruction understanding refers to the process by which an embodied agent interprets and comprehends user input, which can come in various forms, such as text, speech, or gestures. This stage is critical for ensuring that the agent performs the desired tasks correctly, without misinterpretation or miscommunication. The agent's ability to correctly interpret ambiguous or incomplete instructions directly impacts its performance, and any errors could lead to unreliable task execution. Blindly following malicious instructions from users can lead to harmful actions, causing harm to humans and damaging the environment. Ensuring that the agent interprets instructions in a manner that aligns with ethical guidelines is also essential to prevent undesirable outcomes, such as biased or unfair actions.

Environment perception involves the agent's ability to gather and process information from its surroundings using various sensors, such as cameras, LiDAR, or tactile sensors. This stage enables the

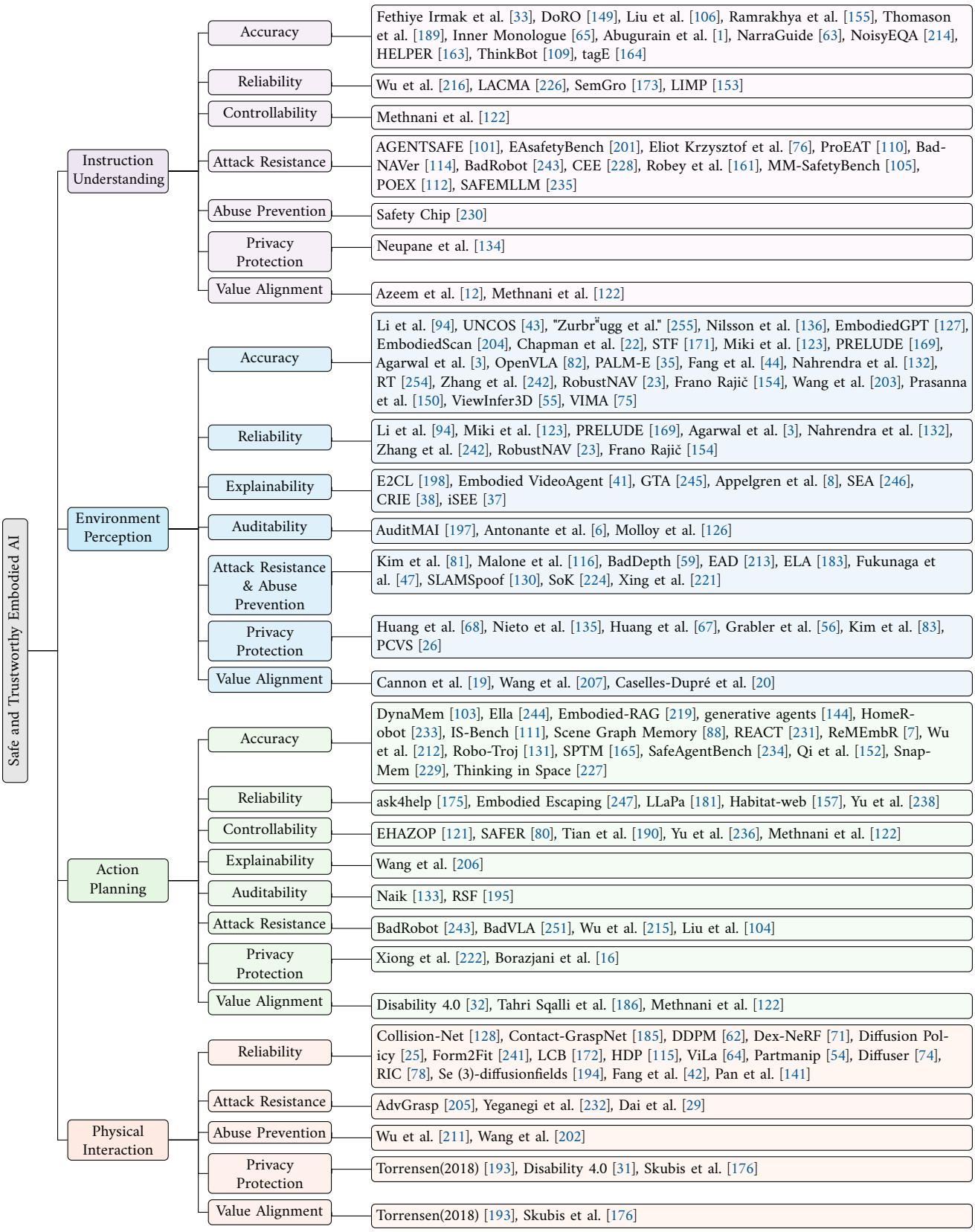

Figure 6 A taxonomy of literature on Safe and Trustworthy Embodied AI.

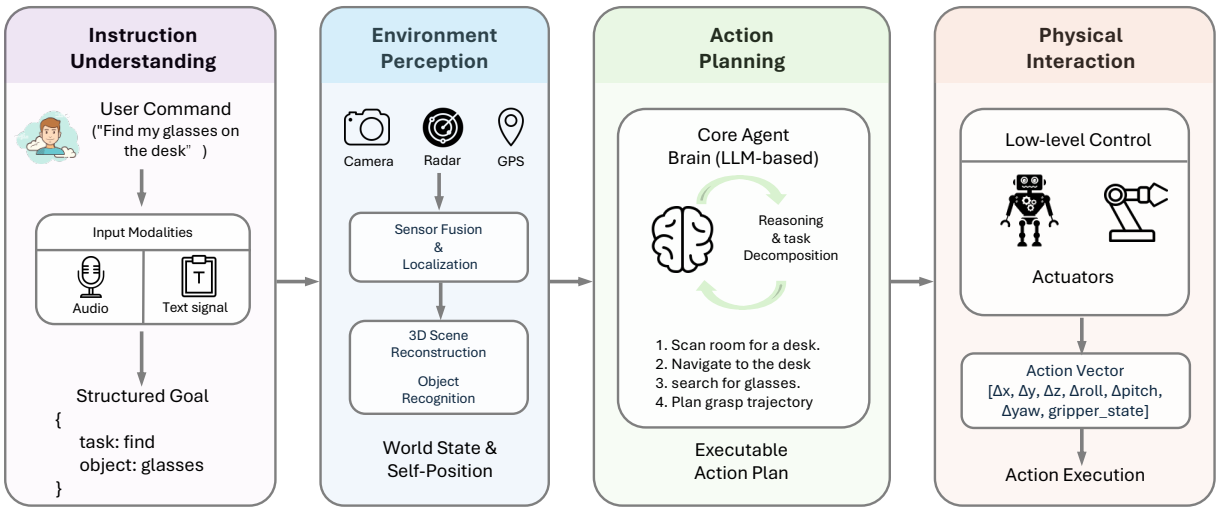

Figure 7 Embodied AI Agent Workflow. The diagram illustrates the four-stage process that an embodied agent follows to complete a task based on a user command. (1) **Instruction Understanding**: A natural language command (e.g., "Find my glasses on the desk.") is parsed into a structured, machine-readable goal. (2) **Environment Perception**: The agent uses sensors like cameras and radar to build a 3D representation of its surroundings and determine its own position. (3) **Action Planning**: A core brain, often based on a Large Language Model (LLM), performs reasoning and task decomposition to create a multi-step executable plan. (4) **Physical Interaction**: The plan is translated into low-level control signals (an action vector) that drive the actuators to execute the task in the real world.

agent to understand the physical context in which it operates, such as detecting obstacles or identifying objects of interest. Perception errors can lead to system failures, such as the agent misinterpreting its environment and taking inappropriate actions based on incorrect data.

Action planning refers to the decision-making process in which the agent determines the sequence of actions needed to achieve a given goal, based on its understanding of the environment and the instructions it received. This step involves complex reasoning and decision-making algorithms that take into account various factors, such as the current environment, the agent's capabilities, and task constraints.

Physical interaction involves the agent executing its planned actions in the real world, such as moving objects, interacting with humans, or navigating through an environment. This stage bridges the gap between decision-making and real-world execution, and it requires the agent to have precise control over its physical movements.

4.2 Instruction Understanding

Instruction understanding is the core interface between an embodied agent and a user, translating human natural language into machine-executable intent. The trustworthiness and security of this process are therefore critical to the entire system's performance. Misinterpretation of an instruction can directly cause downstream planning and execution to fail, while malicious attacks can induce the agent to perform hazardous or illicit actions. In this section, we focus on trustworthy and secure instruction understanding by examining existing research and highlighting representative works across

several key dimensions: accuracy, reliability, attack resistance, abuse prevention, privacy protection, and value alignment.

4.2.1 Accuracy of Instruction Understanding

Interactive Instruction or Dialog. Before the large language model technologies are applied to the field of embodied intelligence, some methods are devoted to realizing interactive instructions. Some methods [50, 189] use decomposed modules to enable the agent to actively ask questions or conduct interactive dialogues. With large language model, some works [163, 65] also allow the agent to ask. This can enrich the single instruction into a whole dialogue, which improves the agent's understanding of user's needs. These methods improve instruction understanding by designing specialized models, thereby improving the success rate of tasks and further improving the accuracy of embodied agents.

Disambiguation. With the widespread adoption of large language models (LLMs) and multi-modal large language models (MLLMs), many challenges in interactive instructions or multi-turn dialogues have been effectively addressed. However, ambiguity in instruction has long remained a critical problem for embodied accuracy. Some methods [106, 149, 109, 164, 214] solve the problem of ambiguous natural language instructions by designing models or specialized data structures. Other methods [33, 155, 63] focus on eliminating ambiguity or instruction ambiguity based on LLM or MLLM. These disambiguation methods enable embodied agents to interpret and execute instructions more accurately by proactively clarifying intent and grounding instructions in context.

Instruction with User Preference. Besides Disambiguation, some studies also focus on user preference to enhance accuracy. A method [1] integrates disambiguation and user preferences into large language models to translate natural language navigation requirements into a safe, high-confidence path planning process.

4.2.2 Reliability of Instruction Understanding

Embodied Instruction Following. Recent works tackle the generalization challenge in embodied instruction following from two complementary perspectives: semantic consistency and environmental adaptability. Works following semantic consistency [226] focus on bridging the semantic gap between high-level language instructions and low-level executable actions. By aligning internal representations with linguistic cues and introducing intermediate semantic abstractions, agents become less reliant on environment-specific visual shortcuts and more faithful to the given instructions. Works following environmental adaptability [216, 173, 153] generalize the instruction understanding ability across diverse or previously unseen environments. Methods in this line dynamically decompose or re-compose high-level plans into low-level actions that respect the current scene affordances, and continuously build or refine scene representations during exploration. Together, these directions enable embodied agents to reliably interpret and execute instructions regardless of domain shifts or prior knowledge about the environment.

4.2.3 Attack Resistance of Instruction Understanding

Attack. Existing research has extensively explored adversarial instruction design to expose deficiencies in the instruction-understanding capabilities of current multimodal large language models (MLLMs). These attack strategies can be broadly categorized into two classes: direct attacks and jailbreak

attacks. Direct Attacks reformulate benign instructions into explicitly malicious ones whose harmful intent is immediately evident. Prior studies [101, 201, 114, 243, 105, 112] have proposed a wide spectrum of harmful instructions and taxonomized them according to the target of harm (e.g., human, environment, or agent itself) or the type of prohibited action (e.g., privacy invasion, physical harm, or misinformation). Jailbreak Attacks craft carefully designed prompts that exploit latent vulnerabilities in the model's safety alignment. While modern MLLMs are often equipped with refusal mechanisms that decline obviously malicious requests, jailbreak attacks seek to circumvent these safeguards and elicit disallowed behaviors. Some works employ hand-crafted prompt templates to achieve this circumvention [114, 243]. Others train a dedicated *Threat Model* to generate adversarial suffixes that systematically evade the model's safety filters [161, 76, 112]. These developments highlight the urgent need to address both direct and jailbreak attacks to preserve the reliability and safety of embodied agents deployed in real-world environments.

Defence. To counter instruction-level attacks, the community has proposed a range of defense mechanisms that can be broadly grouped into introducing extra modules and safety alignment. Extra modules augment the MLLM with auxiliary safety modules that screens incoming instructions before they reach the core model. Some instantiations supply the safety module with carefully designed safety prompts that explicitly instruct the MLLM to reject harmful requests [105, 112, 235]. Others train a dedicated policy model to perform binary or graded safety classification on the raw instruction, ensuring only verified-safe inputs are processed further [112]. While effective, these approaches introduce non-negligible inference-time overhead. Safety alignment seeks to embed robustness directly within the MLLM itself, avoiding additional latency. A first line of work operates on model structure: learnable safety adapters or gated sub-networks are inserted into the transformer stack, allowing the model to up-regulate its own safeguard mechanism when suspicious patterns are detected [228]. A second line leverages adversarial training. By curating datasets of risky instructions and pairing them with appropriate safe responses, researchers design specialized loss functions—e.g., contrastive safety losses or gradient-ascent regularization—that explicitly teach the model to separate harmful from benign instructions in latent space [110]. Together, these complementary strategies advance the goal of delivering both safe and efficient embodied agents for real-world deployment.

4.2.4 Abuse Prevention of Instruction Understanding

Instruction abuse poses a subtle yet serious threat to the safe deployment of embodied agents. While prior studies have explored abuse prevention in general LLM contexts [200, 9, 239], these efforts largely focus on text-only or web-based interactions. In embodied domain, Safety-Chip [230] propose a framework specifically designed for situated, embodied decision-making, where agents must interpret instructions not only for literal correctness but also for contextual and social appropriateness. This role and context aware filtering introduces a nuanced safety layer that goes beyond traditional keyword-based refusal, helping embodied agents detect and resist misuse even when user intent is subtly disguised.

4.2.5 Privacy Protection and Value Alignment of Instruction Understanding

User instructions are not sterile commands; they are often imbued with sensitive personal data and implicit value judgments, which the robot must navigate responsibly.

Privacy in Instructions. Instructions given to an embodied agent can inherently contain sensitive personal information. For example, a command like, "bring my heart medication from the master bedroom" reveals health status, personal habits, and home layout. As highlighted in broader surveys of AI-Robotics security, handling such data is not just a technical challenge but also a legal and ethical one, falling under the purview of data protection regulations like GDPR [134]. The robot's instruction understanding module must therefore be designed not only to parse commands but also to recognize and protect the privacy of the data embedded within them, preventing unauthorized storage, inference, or transmission.

Value Judgments and Discrimination. A more subtle but equally critical challenge arises when instructions require the robot to make value judgments about people. Research shows that even top-tier LLMs, when integrated into robots, produce systematically biased and discriminatory outcomes based on protected identity characteristics mentioned in a prompt [12]. For instance, when tasked with assessing trustworthiness or assigning tasks, the models may exhibit prejudice against individuals described with certain disabilities, nationalities, or genders. This moves the problem of algorithmic bias from screen-based text output to physical, real-world actions.

The Open-Vocabulary Dilemma. The core of this issue is exacerbated by the "open-vocabulary can-of-worms" [12]. Users can unintentionally introduce biased or sensitive descriptors into their natural language commands. The robot must then interpret these ambiguous, value-laden instructions. This creates a critical need for transparency and explainability, as conceptualized in frameworks for trustworthy AI [122]. It is no longer sufficient for the robot to simply execute a command; to be considered trustworthy, it must be able to articulate *why* it interprets an instruction in a certain way, especially when it involves sensitive information or potential social bias. Without robust mechanisms for privacy-preserving understanding and fair value alignment, deploying these systems in uncontrolled human environments remains a significant risk.

4.2.6 Remaining Principles for Instruction Understanding

We deliberately exclude Controllability, Explainability, Auditability, and Identifiability from the scope of instruction-understanding.

- **Controllability** is enforced downstream, not during parsing.
- **Explainability** is an interaction service: the agent justifies its interpretation only when the user asks, not while first decoding the utterance.
- **Auditability** is infeasible at this stage; the high-dimensional, transient, and privacy-sensitive transformer states cannot be logged in real time without crippling latency or leaking personal data.
- **Identifiability** is optional for consumer products; requiring cryptographic or biometric binding for every spoken command adds friction and fails in noisy, multi-user environments.

4.3 Environment Perception

Environmental perception is the foundational stage in real-world applications of embodied systems, so its trustworthiness and security are critical to the whole system. Inaccurate or unreliable perception can compromise downstream decisions, while successful attacks on perception often lead to misinterpretation and system failure. In this section, we focus on trustworthy and secure embodied perception

by examining each core dimension for existing research and highlighting representative works. We will introduce the methods for trustworthy and secure embodied perception from improving accuracy and reliability, explainability, auditability, attack resistance, abuse prevention, privacy protection, and value alignment.

4.3.1 Accuracy and Reliability of Environment Perception

Perception accuracy and reliability mainly encompass two related aspects: intrinsic accuracy under nominal conditions and robustness under challenging or adverse circumstances. Enhancements in embodied perception accuracy, such as improvements in 2D segmentation [136, 255, 43, 150], object detection [203, 44, 87, 22, 171], and 3D grounding [55, 204], directly strengthen the trustworthiness of the perception module. When integrated into embodied systems, these accuracy gains translate into improved overall system accuracy. Some works [75, 35, 254, 127, 82] combine perception modules with pretrained vision–language models and fine-tune them on embodied sensor streams, thereby enhancing both the accuracy and reliability of perception in embodied systems. In the specialized field of trustworthy and secure embodied perception, research efforts predominantly focus on perception robustness in complex scenarios, including studies on robotic locomotion [123, 3, 169, 132] and navigation [23, 154, 242, 94] under challenging real-world conditions.

4.3.2 Explainability of Environment Perception

Perception explainability refers to render the sensory processing of an agent transparent by producing interpretable justifications for raw perceptual outputs. Some work [8, 245, 198, 143] addresses perception errors by allowing agents to explain their incorrect predictions and receive feedback from a teacher or expert, thus making the perception and learning process more interpretable and guided. Some studies [246, 38] focus on self-explanatory methods for embodied perception through generating natural language and other forms of explanation to enhance perception explainability. Other studies [37, 41] attempt to enhance perception explainability by visualizing and analyzing evolving internal models of agents. Overall, these approaches enhance transparency in embodied perception by providing explanations through error-driven feedback, self-generated captions, and analysis of internal representations, clarifying and guiding how perception informs action.

4.3.3 Auditability of Environment Perception

Perception auditability involves maintaining traceable logs and diagnostic information for each perception module at runtime, enabling rapid fault localization, adversarial security analysis, and post-hoc review. For example, the PerSyS framework [6] uses temporal diagnostic graphs to monitor the consistency of heterogeneous perception components and automatically detect failures, ensuring audit-trace completeness and module-level diagnosability; Another work [126] applies HAZOP/Guide-Word techniques to structure safety assessments of perception systems, establishing formal audit checkpoints for each perceptual interface; and the AuditMAI infrastructure [197] introduces automatic audit-trace recording, configurable alerting, and decision-rollback interfaces, demonstrating how auditability can be embedded as a core design property of perception systems.

4.3.4 Attack Resistance and Abuse Prevention of Environment Perception

Environmental perception is one of the initial stages through which an embodied system acquires external information, making research on its attack resistance critically important. Many studies focus on perception attack resistance related to LiDAR [47, 183, 81]. The attack and defensive research on LiDAR includes physical attack methods that tamper with map data through one-step laser deception attacks, as well as evaluations of LiDAR vulnerabilities. For example, a work [130] proposes the SMVS index to determine the optimal attack location, maximizing interference with point cloud matching and correct positioning. Besides, environmental perception encompasses diverse sensing techniques [221], and recent studies focus on attacks and defense on depth estimation [59] and visual positioning tasks [116]. Certain perception-level defense efforts can also be viewed through the lens of abuse prevention [224, 213], which aims to detect and reject manipulated or malicious sensor inputs before they corrupt downstream reasoning and acting.

4.3.5 Privacy Protection of Environment Perception

Studies on Perception Privacy Protection are dedicated to protecting privacy in the perception stage of embodied systems, thus preventing privacy leaks from the initial root. Some work [67, 83, 68] studies privacy protection in embodied perception through different modalities and low resolution at the image level, while a recent work [26] focuses on privacy protection for visual observations from live video streams. Some other works [135, 56] study privacy protection in the perception stage on specific downstream tasks, such as medical care, nursing, etc.

4.3.6 Value Alignment of Environment Perception

Perception-based value alignment mainly centers on how agents use multi-modal perception to understand and adapt to human or environmental value demands. For example, a work [20] studies consistency in the perception-action loop to extract task-relevant features. More studies on value alignment in the perception stage are usually included in research on value alignment in the embodied system. A work [207] quickly infers individual preferences from user interaction data for dynamic value adaptation; another work [19] explores that the perception system itself is the entry point for value embedding from phenomenological and ontological perspectives.

4.3.7 Remaining Principles for Environment Perception

At the perception stage, among the ten core principles of safe and trustworthy embodied AI, controllability and identifiability are not discussed in detail, as they are less relevant to the perception part of the embodied system pipeline. Our focus here is on deterministic perception, which involves interpreting raw sensor inputs in a stable and predictable manner. In contrast, controllability and identifiability are principles that typically apply to generative components, such as decision-making, planning, and action generation, where the system must be able to explain or constrain its possible outputs based on user intent or external feedback. Since perception primarily involves recognizing and interpreting external information rather than generating responses, current research has devoted relatively little attention to applying these principles at this stage. As a result, controllability and identifiability remain underexplored in the context of perception, though they may become more relevant when perception tightly interacts with adaptive or generative mechanisms in the future.

4.4 Behavior Planning

Behavior planning is a core component of embodied intelligence systems, responsible for task decomposition, execution sequencing, interaction strategies, and decision making. As these modules directly govern agent behavior, any failure, misalignment, or adversarial manipulation can lead to critical safety incidents. In this section, we examine key safety and trustworthiness challenges across various categories of planning tasks, organized by the core principles of trustworthy and safe EAI.

4.4.1 Accuracy of Behavior Planning

Accuracy in planning requires a correct understanding of the environment and user intent, ensuring that actions precisely match task goals. A primary challenge is mitigating model "hallucinations"—plausible-sounding plans unsupported by reality, which can be dangerous in embodied contexts. To this end, Qi et al. [152] propose an Embodied Knowledge Graph (EKG) to verify physical plausibility. Similarly, Yao et al. [231] introduce ReAct, a framework that combines reasoning and acting to validate decisions during planning, while Wu et al. [212] present Reinforced Reasoning, which leverages reinforcement learning to guide more grounded planning.

Beyond grounding plans in physical reality, accuracy also depends on an agent's memory, which provides the necessary context to ensure actions are relevant and correct over long horizons. Inaccurate or incomplete memory leads to flawed plans. Early work, such as Semi-Parametric Topological Memory (SPTM) [165], improves navigation accuracy by constructing a graph of observations to enable structured planning. More advanced methods explicitly build and maintain memory representations for a more accurate world model. For example, Active Neural SLAM [21] creates a geometric map that allows a planner to make more precise navigation decisions. To handle dynamic environments, which pose a significant threat to accuracy, systems like Scene Graph Memory [88] and DynaMem [103] update their memory of object states and locations in real-time, preventing the agent from acting on outdated information. Similarly, KARMA [208] employs a dual long- and short-term memory, including a 3D scene graph, to provide planners with a comprehensive and current understanding of the environment, thus improving the accuracy of complex, multi-step tasks.

The accuracy of interpreting user intent also heavily relies on memory. For situated instructions, an episodic buffer is crucial for remembering deferred or conditional tasks, ensuring all components of a command are accurately executed [125]. For social agents, lifelong memory systems like the one in Ella [244], which combines semantic and episodic knowledge, enable more accurate and contextually appropriate social interactions by recalling past events and relationships. Even simulated agents exhibit more accurate, believable behavior when they can reflect on an episodic memory of their experiences [144]. To support complex queries, retrieval-augmented systems build structured memories for precise information access. ReMEmbR [7] uses a vector database of spatio-temporal experiences to accurately answer questions about long-horizon robot activities. Embodied-RAG [219] builds a hierarchical semantic forest, enabling accurate retrieval at multiple levels of granularity. Other approaches focus on the memory representation itself; SnapMem [229] uses a diverse set of visual snapshots as a compact memory for accurate scene reasoning, while other research shows that prompting MLLMs to generate explicit cognitive maps improves the accuracy of their spatial awareness [227]. These works collectively show that a robust and accurate memory is not just a storage mechanism but a fundamental prerequisite for accurate planning and decision-making in complex,

long-running tasks like those found in open-vocabulary mobile manipulation [233].

4.4.2 Reliability of Behavior Planning

Reliability ensures stable and predictable agent performance, especially in complex or unforeseen situations. A key aspect is the agent's ability to recognize its own limitations and degrade gracefully. To this end, some models incorporate self-uncertainty estimates. Yu et al. [238] utilize transparency and credibility of embodied agent planning are improved through confidence assessment, while Singh et al. [175] suggest agents request human help in low-confidence settings. Other work focuses on improving policy robustness; for example, Sun et al. [181] propose LLaPa, which incorporates counterfactual awareness to reduce policy brittleness. Failures in reliability often manifest as generalization gaps in novel environments, as demonstrated in path learning from human demonstrations [157], or as challenges in adapting to constrained spaces [247].

4.4.3 Controllability of Behavior Planning

Controllability ensures that risks are foreseeable and manageable, with mechanisms to prevent harm and allow for human intervention. One approach is to design architectures with explicit safety modules. The SAFER framework [80], for instance, uses a dedicated LLM for safety verification. The need for human oversight is also critical, especially in systems with variable autonomy, to manage trust and governance [122]. In human-robot collaboration, Tian et al. [190] improve planning trustworthiness by enabling robots to adaptively balance safety and efficiency through confidence-aware models of human behavior, while Yu et al. [236] enhance planning trustworthiness by enabling human-robot collaboration systems to synthesize optimal, trust-aware policies under uncertainty using temporal logic and POMDP-based reasoning. Failures in controllability can lead to direct harm, such as collisions caused by miscoordination in social navigation [178]. To proactively identify such risks, methods like EHAZOP offer a tailored hazard analysis process for robot safety [121].

4.4.4 Explainability of Behavior Planning

Opening the "black box" of decision-making is crucial for building trust and enabling meaningful oversight. Research in this area explores methods for making agent behavior understandable to humans. For example, Wang et al. [206] explore explainable learning from demonstration (LfD) as a means to increase human trust in the agent's actions.

4.4.5 Auditability of Behavior Planning

Auditability provides the foundation for accountability by ensuring that an agent's decision-making process is traceable and its actions are verifiable. This is particularly critical in high-stakes applications, where Naik. [133] highlights the challenges in attributing responsibility for failures. To address this, standardized frameworks are being developed; for instance, Adebayo. [195] introduces the Robot Security Framework (RSF), which offers a standardized safety assessment.

4.4.6 Attack Resistance of Behavior Planning

Protecting planning and decision-making systems from malicious manipulation is paramount for safety. Research has identified several vulnerabilities and proposed corresponding defenses. Backdoor attacks are a significant threat, where hidden triggers can exploit planners. Nahian et al. [131] show how LLM-based planners can be exploited, while Zhou et al. [251] proposed a backdoor attack method based on Objective-Decoupled Optimization, which exposes the backdoor vulnerabilities of VLA models for the first time. Other work explores vulnerabilities in neural path planning and proposes gradient-based defenses [222]. Attacks can also occur at the instruction level, such as the physical-world jailbreaks of embodied agents investigated by Zhang et al. [243]. Furthermore, attackers can manipulate the physical environment itself, e.g., by placing objects to degrade motion planners like A* or RRT [215]. Consequently, assessing the adversarial robustness of decision layers in LLM-based embodied models is an active area of research [104].

4.4.7 Privacy Protection of Behavior Planning

As embodied agents operate in personal spaces and handle sensitive data, protecting user privacy is a fundamental requirement. Techniques that enable learning without centralizing raw data are crucial. In this context, Borazjani et al. [16] propose Federated Foundation Models (FFM) to enable safe, adaptive, and personalized embodied AI by combining foundation models with the privacy-preserving, distributed, and user-personalized learning capabilities of federated learning.

4.4.8 Value Alignment of Behavior Planning

Ensuring that an agent's behavior aligns with human ethics and societal norms is a critical challenge. This requires a proactive approach to design. Tahri Sqalli et al. [186] advocate for integrating ethical and inclusive design principles early in the development process, particularly for applications like medical training. This perspective is echoed in specific domains, where researchers examine EAI in assistive robotics from a bioethical standpoint to ensure that technology serves users equitably and responsibly [32].

4.5 Physical Interaction

Physical interaction is fundamental to embodied intelligence systems, enabling precise motion execution, control, environmental manipulation, and adaptive responses to dynamic scenarios. As these modules directly govern an agent's physical behavior, any instability, misalignment, or adversarial interference may result in safety hazards or operational failures. In this section, we explore key safety and trustworthiness challenges in physical control and interaction, focusing on some key components like robustness and secure human-agent collaboration guided by the principles of trustworthy and safe AI.

4.5.1 Reliability of Physical Interaction

In the context of embodied intelligence, reliability in physical interaction has evolved beyond merely maximizing task success rates into a multi-layered, dynamic research frontier. We posit that the paradigm for physical reliability can be deconstructed into three interconnected and progressive

stages: basic execution reliability, which confronts the physical world's uncertainty; behavioral strategy reliability, which masters the multimodality of task solution spaces; and long-horizon task reliability, which ensures the composition and generalization of complex skills. This evolutionary path reflects a profound shift from solving the foundational problem of "whether it can be done" to tackling the advanced challenge of "how to do it more flexibly and intelligently."

Basic Execution Reliability. At this stage, the primary objective is to overcome perceptual noise and physical disturbances to ensure the precise completion of atomic operations, such as grasping and placing. A significant body of work has focused on mitigating execution failures caused by incomplete or inaccurate sensory information. For instance, grasping in cluttered environments, where objects are occluded, requires robust collision avoidance, as addressed by methods like Contact-GraspNet [185] and CollisionNet [128]. Similarly, manipulating transparent objects, which poses a severe challenge to depth sensors, has spurred the development of novel perception techniques. These range from specialized depth completion networks [42] to generative approaches that reconstruct scenes using Neural Radiance Fields (NeRFs) [71] or even leverage diffusion models for view synthesis and inpainting [78]. These methods tightly couple perception and execution, where improvements in perception enhance the determinism of action. By ensuring that each action is stable and well-defined, they establish a solid basis for physical reliability.

Behavioral Strategy Reliability. As basic execution capabilities mature, the research focus transitions to this stage, which centers on acknowledging and effectively modeling the multi-modal nature of a task's solution space. In the real world, a single task often admits multiple valid solutions; for example, a tool can be gripped in various ways, and a path can have multiple obstacle-avoidance options. A truly reliable agent must be capable of generating and selecting from a diverse set of effective strategies, rather than relying on a single, rigid behavior pattern. This shift has catalyzed the rise of generative policy learning methods, most notably those based on diffusion models [62]. By learning to denoise from a simple noise distribution to a valid action, these models can capture the entire complex distribution of the action space. Seminal works like Diffusion Policy [25] and Diffuser [74] have demonstrated that visuomotor policies learned via diffusion can effectively handle the behavioral multimodality inherent in imitation learning datasets. This paradigm extends to complex, non-Euclidean action spaces, such as generating diverse 6-DoF grasp poses on the SE(3) manifold [194]. This capability not only makes policies more generalizable to slight variations in initial conditions but also provides a foundation for handling more complex constraints and preferences, thereby greatly enhancing strategic flexibility and robustness.

Long-Horizon Task Reliability. The frontier of physical reliability is advancing toward this stage. Real-world activities, such as tidying a room or preparing a meal, are inherently long-horizon tasks that require the logical composition of multiple skills. At this level, reliability is no longer about the success of an individual action but the successful completion of the entire task flow. This demands that an agent possess hierarchical planning capabilities and the ability to achieve compositional generalization of its skills. For instance, tasks like kit assembly [241] or manipulating complex articulated objects [54] are quintessential examples of this challenge. Recent studies begin to investigate hierarchical policies, where a high-level planner generates a sequence of subgoals and a low-level policy, often implemented with diffusion models, is responsible for executing each of them [115, 99]. We argue that reliability at this stage is "semantic", requiring the agent not only to execute actions precisely but also to deeply understand the task's structure and goal. This is increasingly being

addressed by leveraging the reasoning and planning capabilities of Large Language Models (LLMs) and Vision-Language Models (VLMs) to guide low-level motion policies [64, 172, 141], bridging the gap between high-level instructions and low-level physical execution.

4.5.2 Controllability and Explainability of Physical Interaction

The process by which embodied agents transform high-level planning into actions in the physical world fundamentally relies on the precise manipulation of physical variables. By computing joint displacements and rotations at each moment, robots are able to perform a wide range of tasks in the physical world. Therefore, how to control and interpret the changes in these variables is a crucial topic in the interaction between embodied agents and their physical environment. This topic is extensively addressed in the field of control theory. As a system, the robot's ability to precisely and rapidly control its internal variables has a long history of research. As early as 1960, Kalman proposed the use of the Kalman rank condition for determining controllability and observability in his paper [77] . Since then, numerous studies have investigated the controllability of robots. For example, the paper [30] introduced the concept of the controllability Gramian, providing criteria for analyzing the controllability of dynamic linear systems. These studies not only ensure that the actions of embodied agents can be accurately controlled during physical movement, but also offer explanations for the execution of each action.

4.5.3 Attack Resistance of Physical Interaction

The robustness against physical attacks at the interaction level primarily focuses on ensuring robots can maintain normal functionality when subjected to external physical disturbances. For locomotion tasks, current research mainly addresses how to preserve robot balance and stability under external force interference. The study [232] introduces a hybrid framework that integrates trajectory optimization with Bayesian optimization to improve the robustness of humanoid locomotion. By tuning cost weights using data from full-body simulations, the method enables the generation of motions that remain stable under various disturbances and uncertainties. Another work [29] introduces a robust anti-disturbance framework based on multi-domain hybrid systems and reduced-order model predictive control, allowing robots to regain balance without falling even when subjected to pushes as strong as 130N. For manipulation tasks, [205] presents an adversarial attack method that interferes with robotic arm grasping tasks by altering objects' geometric shapes.

4.5.4 Abuse Prevention of Physical Interaction

To prevent the abuse of embodied intelligent agents in physical interactions, it is essential to ensure that each control input correctly influences the system, enabling the state variables of the robotic system to evolve as intended. Current research encompasses both studies on how to attack control inputs and strategies to prevent such attacks. In [202], a control input injection method leveraging reinforcement learning is proposed, causing the agent to deviate from its intended trajectory during motion. Conversely, [211] introduces a secure controller designed to prevent robotic systems from collapsing under attack.

4.5.5 Privacy Protection and Value Alignment of Physical Interaction

The physical embodiment of AI and its interaction with humans in real-world settings introduces complex societal, ethical, and psychological challenges. The deployment of humanoid robots in public-facing sectors like tourism and hospitality, for example, raises critical questions about their psychological and emotional effects on both customers and employees, as well as their influence on cultural practices [176]. The ethical stakes are even higher when these systems interact with vulnerable populations. Bioethical analyses warn against a purely functionalist approach to assistive robots for people with disabilities, stressing that failure to consider the person's intrinsic dignity and complex situational factors could be damaging and discriminatory [31]. Ultimately, these specific interaction challenges reflect a broader, urgent need to proactively embed ethical considerations into the design of all robotics and AI systems to avoid the dystopian futures that prominent thinkers have warned against [193].

4.5.6 Remaining Principles for Physical Interaction

The accuracy of physical interaction can be encompassed by the concept of controllability. To ensure that a robot's physical behavior is controllable, it is essential that its actions are executed precisely in accordance with its intended plan. Similarly, auditability is inherently embedded within explainability. When an error occurs during a physical interaction, tracing the source of the problem inevitably involves inspecting certain physical variables. The logical process of such inspection relies on the foundation of explainability. Therefore, the concept of auditability is inherently contained within explainability.

5 Benchmarks and Evaluation

To systematically evaluate and enhance the trustworthiness of embodied agents, the research community develops a diverse array of benchmarks. As shown in Table 1, these benchmarks move beyond traditional task success metrics to conduct in-depth assessments of agents across multiple dimensions, including interactive understanding, explainability, physical safety, and adversarial robustness.

5.1 Interactive Instruction Understanding.

Early research recognizes that trustworthy agents must not only complete tasks but also comprehend the ambiguity and uncertainty inherent in human instructions. To this end, researchers develop benchmarks aimed at enhancing communicative abilities. DialFRED [50] pioneers the introduction of a dialogue mechanism, allowing agents to actively query the user to resolve ambiguity when information is insufficient. Embodied Multi-Agent Task Planning [106] extends this concept to multi-agent collaboration, requiring multiple agents to work together to parse ambiguous instructions. More recent works have further advanced this direction. NoisyEQA [214] focuses on evaluating and improving an agent's ability to identify and correct various types of noise and inaccuracies in user queries, while ASK-TO-ACT [156] explores how reinforcement learning can enable agents to learn to ask the most effective questions at the most opportune moments, thereby efficiently resolving instructional ambiguity. Collectively, these benchmarks have driven the evolution of agents from passive executors to active communicators.

Benchmark	Date	Application	Task	Scenes	Samples	Observation	Simulator	Data Collection	Metrics	Evaluation
EMAT-P [106]	22.06	Instruction Understanding	Planning	120	107035	RGB, depth, text	AI2-THOR	Adapted from ALFRED	SR, SPL	Ground Truth
SafeBench [223]	22.06	Planning	Auto Drive	8	2352	4D, BEV, Cam, Dir	CARLA	Procedural Generation	Safety, Func., Etiquette	–
DialFRED [50]	22.07	Instruction Understanding	EQA	112	53000	RGB, text	AI2-THOR	Human Annotation (AMT)	SR, SPL	Ground Truth
THOR-EAE [206]	23.10	Planning	VLN	120	840000	RGB, action w/ expl.	AI2-THOR	Programmatic Generation	Accuracy	–
MM-SafetyBench [105]	23.11	Instruction Understanding	EQA	13	5040	Image-text pairs	–	Semi-automated Pipeline	ASR, Refusal Rate	–
SEA [246]	24.04	Perception	Visual Affordance	–	9724	Egocentric/Exocentric img	–	Manual Annotation	Grounding, Top-k Acc	–
EARBench [253]	24.08	Instruction Understanding	Planning	28	2636	Text, Visual	–	AI Gen. (Midjourney, GPT-4o)	TRR, TER	LLM as Judge
EAI-Bench [92]	24.10	Planning	Planning	126	438	Instr., trajectory, visual	BEHAVIOR, VirtualHome	Automatic Pipeline	Logic Score, SR	–
Jailbreak-Robots [161]	24.10	Instruction Understanding	EQA	21	105	Image-text pairs	–	From Existing Datasets	ASR	–
NoisyEQA [214]	24.12	Perception	EQA	–	500	RGB, text	–	Adapted from Benchmarks	Acc, DR, CR	LLM as Judge
Harmful-RLbench [112]	24.12	Instruction Understanding	Planning, EQA	25	262	Image-text pairs	CoppeliaSim	Manual Setup in Sim	TSR, ASR, ESR	–
SafeAgentBench [234]	24.12	Instruction Understanding	Planning	Multiple	750	Text, Visual	AI2-THOR	AI Gen. & Human Filtering	Success Rate	Execution, LLM Judge
HASARD [192]	25.03	Planning	VLN	6	–	RGB (+ depth/seg)	ViZDoom	Agent-Env. Interaction	Reward, Cost, Efficiency	Agent Perf. Analysis
SafePlan-Bench [70]	25.04	Planning	Planning	–	2027	Text desc., env. state	VirtualHome	LLM Generation	SafeR, SuccR, ROUGE-L	Rule-based Detector
EAsafetyBench [201]	25.04	Instruction Understanding	EQA	–	9435	Image-text pairs	–	Semi-automated Pipeline	Acc, F1, FNR, FPR	–
ASK-TO-ACT [156]	25.04	Instruction Understanding	EQA	83	–	RGB, joints, state, instr.	Habitat 3.0	From ReplicaCAD	SR, Efficiency	LLM as Judge
IndustryEQA [96]	25.05	Instruction Understanding	EQA	76	1344	Video, Text	NVIDIA Isaac Sim	Human Control & VLLM Gen	Match, Direct, Reason Scores	LLM as Judge
EMBODYGUARD [177]	25.05	Instruction Understanding	Planning	15	942	NL instr., init. state	iGibson & Human Anno.	LLM Gen.	Recall, Interp., Trans. Metrics	Ground Truth
AGENTSAFE [117]	25.06	Planning	VLA	45	9900	Text, Visual	AI2-THOR	Hybrid Gen. (Human+LLM)	PA, PRR, PSR, ESR	Multi-stage (LLM Judge)
IS-Bench [111]	25.06	Instruction Following	Planning	161	388	Multi-modal	OmniGibson	Adapt & Enhance (LLM+Human)	SR, SSR, SRec, SA	Ground Truth, LLM Judge
VPR-Attack [116]	25.01	Perception	Visual Place Rec.	14	158	Query Image	–	From Existing Datasets	Along-Track Error	Ground Truth
BadDepth [59]	25.05	Perception	SLAM	4	16	RGB image	–	From KITTI Dataset	Depth Error Metrics	Ground Truth
BadVLA [251]	25.05	Planning	VLA	130	6500	RGB, text instr.	LIBERO, SimplerEnv	Template-based from Ego4D	Success Rate	–

Table 1 A comprehensive overview and comparison of recent Embodied AI benchmarks. The benchmarks are systematically categorized based on their publication date, application domain (e.g., Instruction Understanding, Planning), and primary task (e.g., EQA, VLN). We further detail their data characteristics, including the scale of scenes and samples, observation modalities, the simulator used, and the data collection methodology. Finally, the table summarizes their evaluation frameworks, detailing the specific metrics and the validation approach (e.g., comparison against ground truth, or using an LLM as a judge). This detailed summary serves as a reference for researchers to navigate the evolving landscape of embodied agent evaluation.

5.2 Decision Transparency and Explainability

To enable humans to understand and trust agent behavior, another critical line of research focus on creating benchmarks for evaluating and generating decision explanations. Generating Explanations for Embodied Action Decision [206] introduces the first large-scale dataset that requires an agent not only to make a decision (e.g., circumvent an obstacle) but also to generate a natural language explanation of *why* that action is optimal. Similarly, Self-Explainable Affordance Learning [246] combines affordance learning with intent expression, requiring the model to generate an "embodied caption" describing its intended action while predicting interactable regions. By "visualizing" and "verbalizing" the agent's decision-making process, these works significantly enhance system transparency and lay the groundwork for establishing human-robot trust.

5.3 Physical Safety and Risk Awareness

Ensuring physical safety is a core requirement for trustworthy embodied AI, leading to the development of numerous benchmarks for assessing agent risk awareness. These can be broadly divided into two categories. The first category focuses on evaluating the safety of static planning at a symbolic or logical level. Examples include EARBench [253], which assesses the physical risk awareness of foundation models in high-level plan generation; SafePlan-Bench [70] and Subtle Risks, Critical Failures [177], which use symbolic environments (e.g., VirtualHome, PDDL) to diagnose subtle risks in plans; and Embodied Agent Interface [92], a diagnostic benchmark that assesses LLMs' core cognitive abilities in embodied decision-making to precisely identify their failure points. The second category takes a step further by evaluating safety during the interactive simulation of dynamic execution processes. SafeBench [223] concentrates on safety-critical scenarios in autonomous driving; HASARD [192] provides an efficient testbed for vision-based safe reinforcement learning; SafeAgentBench [234] systematically evaluate agent responses to hazardous instructions in AI2-THOR; IS-Bench [111] assesses an agent's ability to handle dynamically emerging risks during interaction in OmniGibson; and IndustryEQA [96] extends safety assessment to high-fidelity industrial settings for the first time. Together, these efforts have built a comprehensive evaluation framework for physical safety, spanning from planning to execution and from general household to domain-specific environments.

5.4 Measuring Robustness Against Adversarial Threats

A truly trustworthy system must be resilient to malicious adversarial attacks. To this end, researchers develop benchmarks to assess agents' abilities of attack resistance from different perspectives. At the high-level instruction layer, MM-SafetyBench [105] reveals vulnerabilities in multimodal models to unsafe responses induced by visual content; Advancing Embodied Agent Security [201] focuses on moderating malicious input instructions; and Jailbreaking LLM-Controlled Robots [161] and POEX [112] pioneers research into "jailbreak" attacks that trick agents into performing physically harmful actions, proposing the "Execution Success Rate" metric to measure real-world physical risk. At the low-level perception layer, research has concentrated on stealthier threats like backdoor attacks. BadDepth [59] is the first to systematically demonstrate backdoor attacks against monocular depth estimation models, causing targets to "disappear" from the depth map; BadVLA [251] targets end-to-end Vision-Language-Action models with backdoors; and Adversarial Attacks and Detection in Visual Place Recognition [116] provides a novel experimental paradigm for evaluating the safety of

visual localization systems under adversarial attack. These benchmarks expose vulnerabilities across different attack surfaces and guide the development of more secure embodied AI systems.

5.5 Evaluation Metrics

A rigorous and multi-faceted evaluation framework is paramount for assessing the trustworthiness and safety of embodied agents. Traditional metrics focusing solely on task completion are insufficient for safety-critical applications. Consequently, recent benchmarks introduce a sophisticated suite of metrics designed to quantify performance across distinct principles of trustworthiness and safety. This section systematically categorizes these evaluation methods into four core areas, providing formal definitions and citing their originating works.

5.5.1 Task Performance and Instruction Grounding

Metrics in this category quantify how well an agent's actions align with the explicit and implicit goals of a given instruction, assessing its core competency and understanding. This extends beyond simple task success to include efficiency and the correct interpretation of ambiguous or noisy user inputs.

A foundational metric is the Success Rate (SR), which provides a binary measure of task goal completion. This is often complemented by the Success weighted by Path Length (SPL), which penalizes inefficient paths, thereby measuring both correctness and efficiency [106, 50]. The SPL is formally defined as

$$\text{SPL} = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{L_i^*}{\max(L_i, L_i^*)}, \tag{1}$$

where N is the total number of episodes, S_i is a binary success indicator for episode i, L_i^* is the length of the optimal path, and L_i is the length of the agent's actual path.

For tasks involving ambiguity, specialized metrics are required. The Ambiguity-Resolution Efficiency Score (ARS) evaluates an agent's ability to ask minimal, relevant questions to successfully complete a task [156]. It is calculated as

$$\text{ARS} = \frac{\mathbb{I}_{\text{success}}}{1 + |\text{q}_{\text{relevant}} - K| + \text{q}_{\text{irrelevant}}}, \tag{2}$$

where $\mathbb{I}_{\text{success}}$ is a success indicator, $\text{q}_{\text{relevant}}$ and $\text{q}_{\text{irrelevant}}$ are the numbers of relevant and irrelevant questions asked, and K is the minimum number of questions required. Similarly, to evaluate robustness against noisy queries, the Noise Detection Rate (DR) and Noise Correction Rate (CR) measure the agent's ability to identify and rectify factual errors in user instructions [214]:

$$\text{DR} = \frac{|A^d|}{|A|} \times 100\%, \quad \text{CR} = \frac{|A^c|}{|A|} \times 100\%, \tag{3}$$

where $|A|$ is the total set of answers, while $|A^d|$ and $|A^c|$ are the subsets of answers that successfully detect or correct the noise, respectively.

5.5.2 Safe and Reliable Plan Execution

Metrics for safe and reliable execution assess the quality of the agent's generated plan, focusing on its physical executability and adherence to safety constraints. A reliable plan must first be executable. The Execution Success Rate (ESR) or Execution Rate (ER) measures the proportion of a plan's steps that can be successfully run by the low-level controller without errors, diagnosing the crucial gap between high-level planning and physical embodiment [117, 234]. Frameworks like *Embodied Agent Interface* [92] provide a fine-grained error analysis, breaking down failures into categories such as "Missing Step" and "Affordance Error" for deeper reliability diagnostics.

More importantly, a reliable plan must be safe. Safety metrics evaluate the agent's capacity to avoid causing harm. The Task Risk Rate (TRR) measures the fraction of plans containing potential physical risks [253]:

$$\text{TRR} = \frac{\sum_{i=1}^{N} I_s(p_i, s_i)}{N}. \tag{4}$$

where $I_s(p_i, s_i)$ is an indicator function that is true if the plan p_i violates safety guidelines s_i. Other benchmarks introduce a suite of safety-centric success rates, such as the overall Safety Rate (SafeR) [70] and the Safe Success Rate (SSR) [111]. A crucial process-oriented metric is the Safety Recall (SRec), which evaluates whether safety-critical actions are performed correctly within the task flow [111]:

$$\text{SRec} = \frac{\sum_{g \in G_{\text{safe}}} \mathbb{I}(g \text{ is triggered} \wedge g \text{ is satisfied})}{\sum_{g \in G_{\text{safe}}} \mathbb{I}(g \text{ is triggered})}. \tag{5}$$

This metric uniquely captures whether safety goals ($g \in G_{\text{safe}}$) are met when they become relevant (i.e., are triggered).

5.5.3 Decision Transparency and Explainability

Explainability metrics assess the quality of natural language explanations generated by an agent to justify its decisions, opening the "black box" of its reasoning process. To evaluate the factual correctness of an explanation's content, the F1-score is commonly used, which is the harmonic mean of Precision (P) and Recall (R) [206]:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{7}$$

$$\text{F1} = 2 \cdot \frac{P \cdot R}{P + R}. \tag{8}$$

For fluency and semantic similarity, standard NLP metrics like BLEU-n [206] and CIDEr [246] are employed. The BLEU score measures n-gram overlap with a brevity penalty (BP):

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{9}$$

where p_n is the modified n-gram precision. The CIDEr score evaluates consensus by computing the average cosine similarity between the TF-IDF vectors (g^n) of the candidate sentence (c_i) and the

reference sentences (s_{ij}):

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{M} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \cdot \|g^n(s_{ij})\|}. \tag{10}$$

5.5.4 Security against Malicious Use and Attacks

Security metrics evaluate an agent's resilience against external threats, encompassing both the explicit rejection of malicious instructions and the defense against subtle adversarial attacks. The first line of defense is abuse prevention, measured by the agent's ability to refuse harmful commands. The primary metrics here are the Rejection Rate (Rej) [234] and the Planning Rejection Rate (PRR) [117, 177], calculated as the fraction of hazardous instructions the agent explicitly refuses to execute:

$$\text{PRR} = \frac{|D_{\text{reject}}|}{|D_{\text{eval}}|}, \tag{11}$$

where $|D_{\text{reject}}|$ is the subset of correctly refused instructions.

When prevention fails, attack defense metrics quantify the impact. The most prevalent metric is the Attack Success Rate (ASR), which measures the frequency with which an attack successfully coerces the agent into performing a harmful action [251, 112]:

$$\text{ASR} = \frac{\text{Number of successful jailbreaks}}{\text{Number of attempted jailbreaks}}. \tag{12}$$

The consequences of a successful attack are measured with task-specific metrics, such as increased Along-Track Error (ATE) in navigation, or catastrophic failure events like Loss of Vehicle (LoV) [116]. For perception-level attacks, the degradation of standard metrics like absolute relative error within the targeted image region can signify a successful backdoor injection [59].

5.6 Future Directions: Towards a Unified and Dynamic Evaluation Framework

While foundational, the current landscape of embodied AI benchmarks reveals a fragmented ecosystem struggling to keep pace with the field's rapid advancements. Most evaluation platforms are static, limited in their coverage of dynamic and long-tail risks, and suffer from a persistent sim-to-real gap. Furthermore, the lack of standardized data representations and metrics makes it difficult to compare agent performance across different embodiments and environments. To foster the development of truly robust and reliable agents, the community must move towards a unified, dynamic, and extensible evaluation framework. This next-generation paradigm would be built on a standardized architecture capable of harmonizing heterogeneous 3D data and a normalized metric library, enabling fair, cross-platform comparisons.

At the core of this future framework will be high-fidelity, interactive, and editable virtual worlds. To bridge the sim-to-real gap, these environments must not only be visually realistic but also physically accurate, powered by advanced physics engines that simulate complex dynamics. Crucially, they must allow users or automated systems to perform real-time, personalized editing—dynamically altering

object properties, introducing unexpected obstacles, or modifying task constraints during execution. This interactivity transforms evaluation from a static checklist into a targeted, adversarial process. To populate these worlds at scale, the framework will leverage generative AI, using techniques like diffusion models and LLMs to automatically synthesize a vast and diverse array of scenes and logically complex tasks, thus overcoming the bottleneck of manual creation.

Ultimately, the most profound shift will be from one-off assessments to a continuous, closed-loop evaluation system. This paradigm views evaluation as an integral part of the agent's development lifecycle, creating a co-evolutionary process. In this "generate-execute-analyze-evolve" loop, the system would automatically generate tasks tailored to an agent's current capabilities, analyze its performance to pinpoint specific failures or weaknesses, and then use that analysis to generate new, targeted challenges. This creates a symbiotic relationship where the agent's capabilities and the platform's evaluation rigor improve together, systematically driving embodied AI towards a higher level of generalizable and trustworthy intelligence.

6 Simulator

Name	Scenes / Rooms	Objects / Cat.	Physics Engine	Scene Source	Customizable	Editable	Action Space	Multi-agent
AI2-THOR [85]	-/120	118/118	Unity	Modeling	✓	I, M	N, F, A	✓
CALVIN [118]	4/-	7/5	Bullet	Modeling	✗	I, M	N, F	✗
CHALET [225]	10/58	330/150	Unity	Modeling	✗*	I, M	N, A	✗
DMC [187]	1/-	4/4	MuJoCo	Modeling	✓	I	F	✗
InternUtopia [199]	100000/-	24957/956	PhysX	Modeling	✗*	I, M	A, N	✗
Gazebo [84]	–	–	Open Dynamics Engine	Modeling	✓	I	F	✓
Genesis [250]	–	–	(Proprietary)	Modeling	✓	I, M	N, A, F	✓
Gibson [217]	572/-	–	Bullet	Scanning	✗	N	N, F	✗
Habitat [166]	–	–	Bullet	Scanning	✓	N	N	✗
Isaac Sim [137]	–	–	PhysX	Modeling	✓	I, M	N, A, F	✓
LabUtopia [95]	100	-/140	PhysX	Modeling	✓	I, M	N, A, F	✗
Meta-World [237]	1/-	80/7	MuJoCo	Modeling	✓	I, M	F	✗
RLBench [73]	1/-	28/28	Bullet	Modeling	✓	I, M	F	✗
SAPIEN [218]	–	2346/-	PhysX	Modeling	✓	I, M	A, F	✗
ThreeDWorld [48]	15/120	112/50	Unity, Flex	Modeling	✓	I, M	A, F	✓
UNREALZOO [249]	100/	–	Unreal	Modeling	✓	I, M	N, A, F	✓
VRKitchen [51]	16/16	–	Unreal	Modeling	✗	I, M	N, A, F	✗
VirtualHome [151]	7/-	-/509	Unity	Modeling	✗	I, M	N, A	✗
airsim [170]	–	–	Unreal	Modeling	✓	N	F	✗
iGibson [91]	15/108	152/5	Bullet	Scanning	✗	I, M	N, F	✗

Table 2 A Comparison of Major Simulators for Embodied AI. Column definitions are as follows. **Scene Source**: Modeling indicates scenes created with 3D software; Scanning indicates scenes reconstructed from real-world scans. **Customizable**: ✓indicates the ability for users to create new scenes; ✗ indicates a lack thereof. An asterisk (*) denotes the ability to recombine existing assets but not create entirely new scenes. **Editable**: I for interactable objects; M for multi-state objects; N for non-editable/static scenes. **Action Space**: N for navigation actions; A for atomic interactions; F for fine-grained force/torque control. **Multi-agent**: ✓indicates support for multiple agents; ✗ indicates no support. A double dash (–) indicates data not specified in the source.

Simulators are essential tools in Embodied AI research, providing scalable, parallelizable, and safe environments for training and evaluating intelligent agents. To develop trustworthy and reliable

agents, two dimensions are of paramount importance: *scene customization* and *environmental editability*. *Scene customization*—the ability to generate or import novel environmental layouts—is fundamental to building trustworthy agents. An agent can be deemed reliable, robust, and predictable only if it is validated across a wide distribution of diverse scenarios, rather than a limited set of pre-defined. This capability allows researchers to rigorously evaluate an agent's generalization limits, assess its capabilities and failure modes, and ensure its behavior remains controllable. From a safety perspective, programmatic scene generation enables the creation of adversarial or edge-case scenarios (e.g., cluttered pathways, unusual object placements) to proactively identify and mitigate potential risks, ensuring the agent operates safely even in unforeseen situations. *Environmental editability*—the extent to which an agent can interact with its environment, ranging from basic physical dynamics to complex object state changes—is equally critical. It directly influences the trustworthiness of an agent by enabling it to learn and perform complex, meaningful tasks. An agent that can accurately manipulate articulated objects or change their intrinsic states (e.g., slicing a fruit) demonstrates greater reliability, and its abilities more closely reflect real-world competence. From a safety perspective, simulating fine-grained interactions allows researchers to evaluate agent behavior, ensuring that actions are identifiable and that associated risks remain controllable. This, in turn, helps prevent unintended harmful outcomes and ensures that the agent's interactions with the environment are both effective and safe. We summarize the key features of prominent simulators in Table 2. This review categorizes prominent simulators along these two dimensions to create a clear landscape of available tools for embodied AI research, as illustrated in Figure 8.

6.1 Pre-defined, Static Scenes

This category includes simulators that offer high-fidelity, realistic scenes but with limited interactivity and no support for creating new environments. Such simulators are primarily suited for navigation and perception tasks in realistic settings. A representative example is the original Gibson environment [217]. It provides a large dataset of scenes reconstructed from real-world 3D scans. While this offers unparalleled visual realism for navigation tasks, the scenes are static and do not support object manipulation. The agent's interaction is limited to navigation and collision with the static mesh of the environment.

6.2 Pre-defined, Interactive Scenes

Simulators in this category provide a fixed set of environments but enrich them with highly interactive objects that can change states. These platforms are excellent for research on long-horizon planning, task decomposition, and complex manipulation skills. This category includes several benchmarks designed for complex household tasks. For example, CALVIN[118] and VirtualHome[151] feature agents performing long-horizon activities involving objects with discrete states (e.g., opening drawers, switching on lights). Others emphasize finer-grained realism; VRKitchen[51] simulates continuous state changes like slicing vegetables, while iGibson 2.0[91] incorporates physical states like temperature and wetness into scanned environments. Similarly, CHALET[225] allows for programmatic object placement within its fixed houses to test generalization across diverse object configurations. To enhance scene diversity, UnrealZoo[249] provides a large collection of 100 interactive worlds, ranging from indoor to urban environments, making it valuable for evaluating generalization across dynamic scenarios. A distinct sub-category focuses on task generalization within a single, static scene.

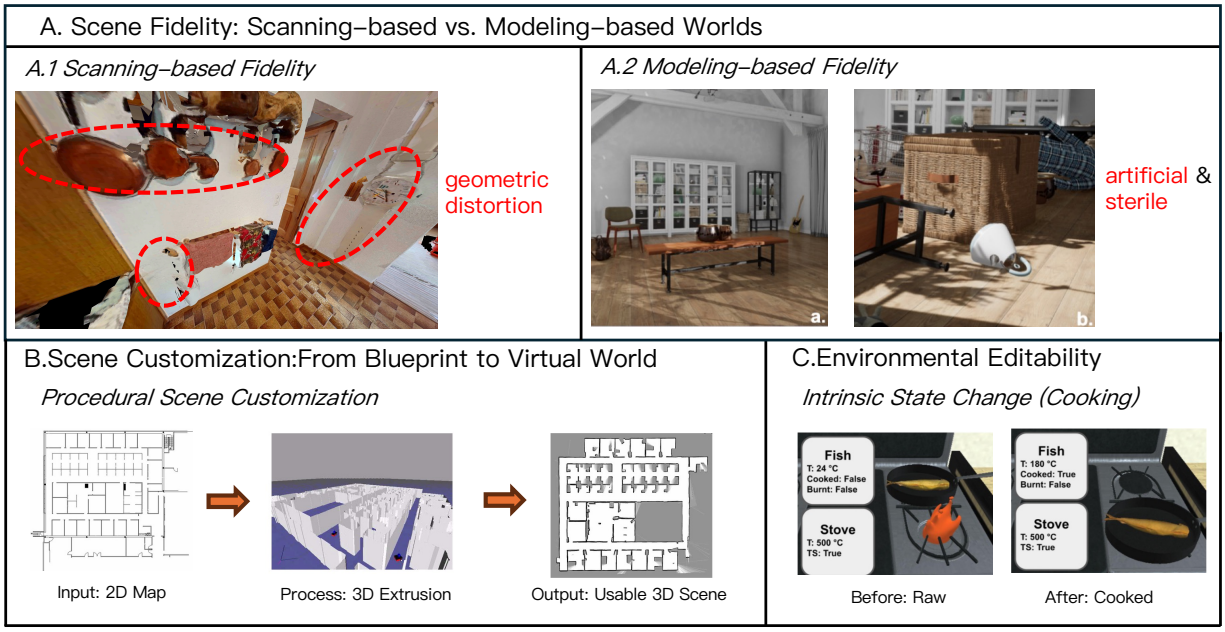

Figure 8 Visualizing Key Dimensions for Evaluating Embodied AI Simulators. This figure illustrates three critical dimensions—scene fidelity, customization, and editability—that are essential for developing trustworthy agents. **(A) Scene Fidelity:** Comparison of scene generation methods. (A.1) A scene from a real-world scan (Habitat [166]) may contain artifacts such as geometric distortions. (A.2) Scenes from a modeling-based engine ThreeDWorld [48] are geometrically accurate but appear artificial and sterile. **(B) Scene Customization:** An example of procedural scene generation from Gazebo [84]. A simple 2D map is programmatically converted into a novel 3D environment, a crucial capability for testing agent generalization across diverse scenarios. **(C) Environmental Editability:** An example of intrinsic state change from iGibson 2.0 [91]. An agent's interaction transforms an object's state from 'Raw' to 'Cooked', enabling the simulation of complex real-world tasks beyond simple navigation and manipulation.

Meta-World[237], for instance, offers 50 distinct manipulation tasks in one tabletop setting with randomized object and goal positions, making it a standard for multi-task and meta-reinforcement learning. Similarly, RLBench[73] features over 100 tasks and a powerful API for creating new ones, positioning it as an ideal testbed for skill acquisition, imitation learning, and few-shot learning. Finally, some platforms serve as foundational frameworks rather than navigable environments. The DeepMind Control Suite [187], built on MuJoCo, offers a rich set of pre-defined control tasks, acting as a standardized tool for benchmarking reinforcement learning algorithms.

6.3 Customizable, Static or Low-Interaction Scenes

This category is characterized by platforms that allow users to create or import new scenes but offer limited physics-based interaction. They are suitable for large-scale navigation experiments across diverse environments but not for manipulation-heavy tasks. Habitat [166] is a key example. It is a high-performance simulator optimized for navigation, allowing users to easily import and use their own 3D scene datasets (e.g., from scans). However, its initial versions lacked physics simulation, rendering all objects static. AirSim [170], built on Unreal Engine, also fits here. It allows users to create or use any environment within Unreal Engine, offering high customization. Its focus, however, is on

vehicle (drone, car) simulation, and it does not support agent-based object manipulation.

6.4 Highly Customizable and Interactive Scenes

This category features simulators that provide powerful APIs for creating diverse new scenes and support rich, physics-based interactions with multi-state objects. These platforms provide a strong foundation for developing general-purpose, robust, and highly capable embodied agents. We can group these platforms into general-purpose engines and task-oriented simulators.

General-Purpose Platforms and Engines: NVIDIA Isaac Sim [137] stands out as a powerful, robotics-focused platform built on PhysX. It offers extensive scene customization through Python APIs and asset importers (URDF, MJCF) and supports complex, multi-state object interactions and large-scale, multi-agent simulations. ThreeDWorld [48] is another versatile platform supporting multi-modal simulation (including audio and soft-body physics) and procedural generation of rich, interactive environments. Genesis [250] aims to be a generative and universal physics engine that can create scenes and tasks from high-level prompts. The open-source Gazebo simulator [84] supports multi-robot simulation with high customizability for models and environments.

Task-Oriented and Extensible Simulators: Many simulators provide both rich default content and strong extension capabilities. AI2-THOR [85], especially with its ProcTHOR extension, excels at the procedural generation of interactive indoor environments where objects can be manipulated in complex ways (e.g., sliced, cooked, opened). SAPIEN [218], with its focus on part-based articulated objects from the PartNet-Mobility dataset, is ideal for tasks requiring a deep understanding of object mechanics. Other simulators extend the frontiers of scale and domain. InternUtopia [199] introduces city-scale environments with thousands of interactive scenes and social, LLM-driven NPCs. Its scene customization is primarily reflected in the ability to modularly combine a massive library of pre-made scenes into new, larger-scale city environments, rather than procedurally generating individual scenes from scratch. LabUtopia [95] focuses on scientific laboratory environments, featuring a procedural scene generator and simulating not just physics but also chemical state changes.

7 Position and Future Directions

Our systematic review in Chapters 4, 5, and 6 reveals a vibrant yet fragmented research landscape. While significant progress has been made in perception, planning, control, and safety evaluation, these efforts often advance in isolated silos. We posit that the path toward truly trustworthy embodied AI necessitates a paradigm shift from optimizing individual components to engineering a holistic, closed-loop system. Inspired by the principles of cybernetics, we argue that trustworthiness is not a feature we can simply "add" to an agent. Instead, it is a quality that emerges from the agent's constant, dynamic interaction with its environment and other beings.

From this perspective, a trustworthy embodied agent should be conceptualized not as a pre-programmed machine, but as an advanced adaptive control system. As illustrated in Figure 9, this system operates within a classic feedback loop. The Environment, which includes the physical world, humans, and other agents, serves as the complex, unpredictable "plant" to be controlled. The Agent itself acts as the

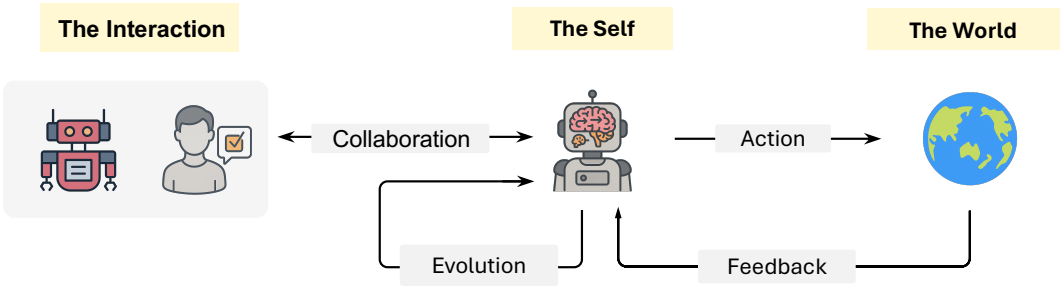

Figure 9 Our proposed cybernetic framework for trustworthy embodied AI, visualized as a closed-loop system built upon three fundamental pillars. **The Self** represents the agent, which acts as an adaptive controller. It performs an **Action** in **The World** (the environment), which provides sensory **Feedback**. This feedback drives the agent's internal **Evolution**, enabling it to learn and adapt. Simultaneously, the agent engages in **Collaboration** through **The Interaction** interface, which encompasses complex interactions with humans and other agents. This virtuous cycle illustrates our position that trustworthiness emerges from the continuous interplay of these three pillars.

adaptive controller. It receives a high-level Task as its reference signal and continuous Perception as its feedback signal.

The core of our position is that the trustworthiness of this entire system hinges on the quality of three fundamental pillars that constitute this loop:

The World: The quality of the feedback signal is determined by the realism of the environment. A trustworthy agent can only be forged through interaction with a high-fidelity world that provides meaningful, realistic consequences for its actions.

The Self: The agent itself must be adaptive. It needs an intrinsic self-evolution mechanism that allows it to learn from the feedback loop, continuously updating its internal model and improving its control strategy over time.

The Interaction: The Sophistication of the Interface. The agent's ability to process inputs and produce effective outputs depends on its coordination architecture. This interface must seamlessly integrate internal brain-body synergy with external multi-agent and human-agent collaboration.

This chapter will now elaborate on each of these pillars, culminating in a unified vision for a research agenda aimed at building and perfecting this closed-loop system for trustworthiness.

7.1 The World: Bridging the Reality Gap with High-Fidelity Scalable Interactive Virtual Environments

In Chapter 4, we see that most existing research on trustworthy embodied AI uses limited real data [59] or handcrafted simulators [85, 166] to address a specific trustworthiness or security issue. Research on the entire process of embodied systems is lacking, indicating that progress in trustworthy embodied AI is severely constrained by the development of the basic capabilities of embodied intelligence. However, in contrast to fields such as computer vision and natural language processing that have access to large-scale general-purpose data, large-scale high-quality data for trustworthy embodied intelligence is extremely scarce since the research on embodied intelligence relies on the interaction between agents

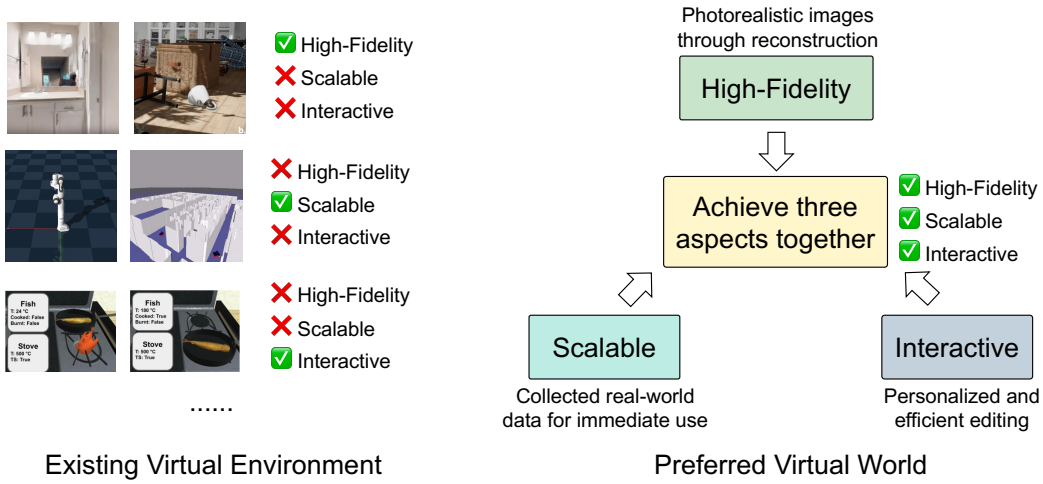

Figure 10 Existing virtual environment and the virtual world we prefer. We argue that **high-fidelity**, **scalable**, and **interactive** are three key aspects of a virtual environment. Existing methods fail to meet all requirements simultaneously, while the preferred virtual world should achieve three aspects together.

and physical environments, and the data forms of different embodied tasks vary greatly. Therefore, researching trustworthy embodied intelligence in virtual environments has become an inevitable choice at present.

Existing simulators [91, 51, 151, 118] lack realism and are difficult to scale up under limited resources, failing to capture the authenticity and scene diversity of real environments. We argue that research on trustworthy embodied intelligence requires constructing high-fidelity, scalable, interactive virtual training environments that bridge the reality gap to the real world, which could significantly facilitate progress toward the entire system of trustworthy embodied intelligence.

The constructed virtual environment should have the following capabilities: (1) Efficient mapping mechanisms among heterogeneous 3D representations in the virtual environment. Different embodied tasks often rely on different forms of 3D representation [54, 71], while existing virtual scenes typically use fixed representations [85, 166, 51, 151, 170, 249], and a single unified representation lacks a feasible solution. Therefore, the virtual environment must support mappings across multiple representations to enable flexible transfer and efficient adaptation of virtual scenes among diverse embodied tasks, thereby improving task generality. (2) High-fidelity and scalable virtual environments. The virtual embodied training environment should directly leverage raw sensor data from real-world environments to construct high-fidelity scenes, and support dynamic expansion mechanisms, such as via the internet or real-time data collection, to increase scene diversity and scale, meeting the requirements of embodied research for varied and large-scale scenarios. (3) Personalized interaction capabilities in the virtual environment. The constructed virtual environment should allow objects to be manipulated via convenient interfaces, such as natural language, enabling efficient interaction, and providing embodied systems with real-time operational feedback and dynamic scene updates.

As shown in Figure 10, these capabilities, including high fidelity, scalability, and interactivity, form an "impossible triangle" in existing virtual-scene construction methods, as current techniques typically address at most one or two of these aspects and cannot satisfy all three simultaneously. If the limitations in balancing realism, generality, and interactivity are addressed, we can construct high-fidelity, scalable,

and interactive virtual embodied environments that support training for diverse embodied tasks and ultimately enable trustworthy embodied system solutions.

7.2 The Self: From Pre-Trained Statues to Self-Evolving Embodied AI

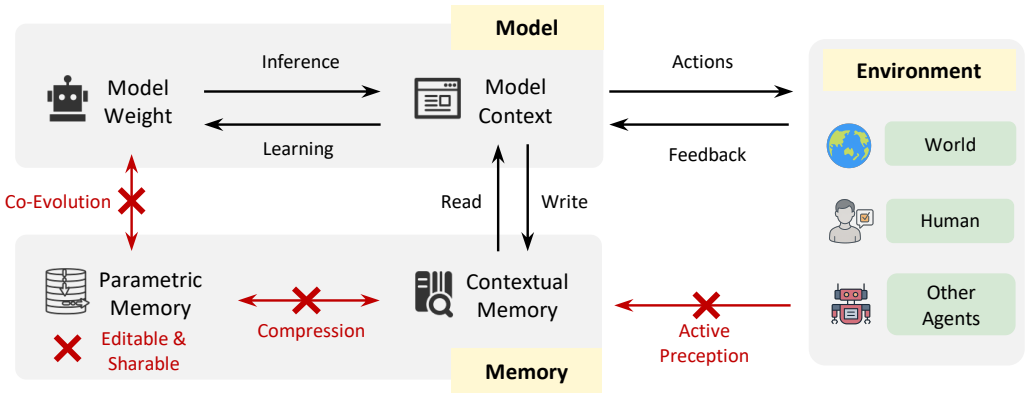

Figure 11 A next generation memory system holds promise as the key component toward EAI evolution. Compared to chatbot agents, embodied agents should incorporate parametrized active memory systems to compress multi-model and life-time scene memories and participate in co-evolution with the model parameters themselves. Black components represent mature technologies (RAG, continual learning, working memory, etc.), while red annotations indicate immature technologies.

In Chapter 4, we saw two different ideas for safety: building it from rules or learning it from data. Both methods today lead to agents that are like "pre-trained statues." Once they are created, their abilities are fixed. This makes them fragile. They cannot adapt to new dangers, and we cannot be sure their safety rules will work in new situations. We argue that trustworthy agents cannot just be "built"; they must be able to evolve. They need their own ways to improve themselves based on their experience in the world.

World experience, comprising perception-action data accumulated by embodied agents during environmental exploration, constitutes the foundation of their knowledge acquisition. Current research predominantly employs external memory systems to structure these experiences into retrievable knowledge representations and episodic memories [57, 14]. However, this approach fundamentally remains an in-context engineering mechanism that, while enhancing agent performance in specific environments and tasks [220], lacks the means to continuously improve agent capabilities through systematic utilization of the memory system. Embodied continual learning [209, 28] enables direct integration of accumulated experiences into model parameters [120]. Nevertheless, this approach faces technical challenges including prohibitive computational costs, training instability, and catastrophic forgetting [18]. More critically, existing continual learning paradigms scarcely constitute self-evolution, instead relying upon external evaluation criteria and training pipelines.

As shown in Figure 11, a self-evolving embodied memory systems must incorporate the following capabilities: (1) **Active Preception**, memory systems should actively perceive the environment and assess trajectories [102, 160] based on historical experience , rather than merely serving as passive LLM tools. (2) **Parametric Memory compression** for life time multi-model memory, enable a comprehensive observation of the "world" [24, 100]. (3) **Shareable and editable** memory mechanisms, with related work including memory interchange protocols and multi-agent memory sharing

frameworks [49, 97]. (4) **Memory-Model co-evolution** mechanisms, parameters should achieve natural alignment and co-evolution within a unified semantic space, rather than relying upon external auxiliary modalities [98, 252, 179].

Ultimately, we can construct a next-generation intrinsic embodied memory system that fully serves trustworthy model capability evolution. This system transcends rigid "retrieval-generation" mechanisms, establishing an iterative pathway from environmental perception → decision execution → self-memory updating → continual learning. This framework integrates the stability advantages of memory systems with the adaptive capabilities of continual learning, providing a theoretical foundation for the long-term autonomous development of embodied intelligence.

7.3 The Interaction: Achieving Seamless Coordination

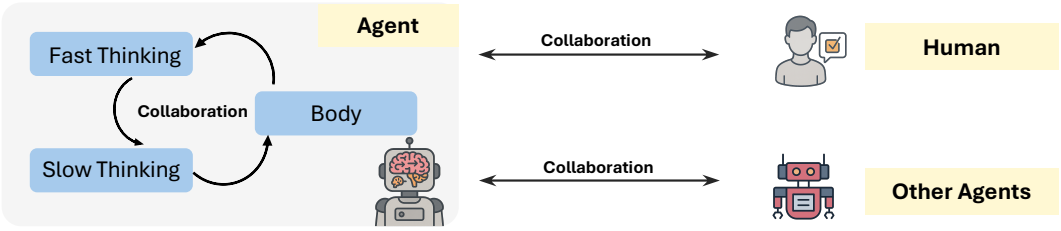

Figure 12 An illustration of the third pillar of seamless coordination: Internal, Multi-Agent, and Human-Agent. The agent's ability to effectively integrate these three interaction channels is fundamental to achieving robust trustworthiness in the physical world. Each channel addresses a unique challenge, from bridging the internal semantic-physical gap to enabling safe collective behavior and intuitive human collaboration.

Many trust failures in embodied intelligence arise not from a single faulty component, but from poor coordination. This friction can occur as a disconnect between the agent's deliberative "brain" and its reactive "body," or as a breakdown in its interactions with humans and other agents.

We believe that trustworthiness is not a property of an agent in isolation but emerges from the quality of its interactions. Therefore, to build truly trustworthy agents, we must develop better architectures and protocols for seamless coordination at three critical levels: internal (brain-body), multi-agent, and human-agent, as illustrated in Figure 12.

Before delving into these three levels, it is noteworthy that a foundational principle of trustworthy interaction, Identifiability, has received little attention. This principle ensures an agent's presence and actions are clearly distinguishable from those of a human to prevent deception. Indeed, our quantitative literature analysis (Figure 5) shows that research in this area is still nascent. While we focus on the three levels of coordination below, we consider Identifiability a cross-cutting concern for all forms of interaction.

Internal Coordination: Brain-Body Synergy via Fast and Slow Thinking The most fundamental challenge in internal coordination is bridging the "semantic-physical safety gap." We posit this gap can be powerfully framed through the lens of Fast and Slow Thinking, which distinguishes between an agent's "Slow Thinking" brain (the deliberative, LLM-based planner) and its "Fast Thinking" body (the reactive, low-level motor controller). Trustworthiness failures often occur when the "slow," semantically-aware plan is incompatible with the "fast," dynamic realities of physical execution. Therefore, the research

frontier is to develop unified architectures that enable a seamless synergy between fast and slow thinking, a direction already being explored in frameworks that learn to make a decision between deliberative and reactive modes [180, 182]. The emergence of powerful, end-to-end Vision-Language-Action (VLA) models like Helix, which fuse high-level reasoning with low-level motor control in a single system, represents a significant step toward this goal [46].

Multi-Agent Coordination: From Individual Safety to Collective Trust Coordination extends beyond the agent's own mind and body to its interactions with other intelligent agents [58]. In shared spaces like warehouses or public roads, the safety of one agent is inextricably linked to the predictability and cooperativeness of others. Therefore, a trustworthy system requires robust protocols for multi-agent coordination. This involves designing rules for negotiation, signaling intent, and decentralized teamwork, ensuring that a group of agents can act as a safe and reliable collective, rather than a chaotic crowd.

Human-Agent Coordination: Beyond Control to Shared Autonomy The ultimate form of external coordination involves interaction with human users. A truly trustworthy agent must be more than a simple tool; it must be a competent collaborator. This requires moving beyond simple remote control to a system of shared decision-making, often referred to as shared autonomy. In such a system, humans provide high-level, goal-oriented direction, while the agent executes the low-level details. Critically, this collaboration must be bidirectional: the agent needs to infer the user's intent from their actions [158], and in turn, it must also be able to clearly communicate its own capabilities, limitations, and uncertainty back to the human partner, ensuring that the human can make informed decisions and build a well-calibrated sense of trust [45].

7.4 A Unified Vision for the Future

The three pillars we have outlined, the World, the Self, and the Interaction, are not independent research avenues but the essential, interdependent components of the cybernetic control system illustrated in Figure 9. The future of trustworthy embodied AI lies not in perfecting any single component in isolation, but in understanding and engineering the synergies of the entire closed loop.

This virtuous cycle represents the process by which trustworthiness is continuously generated and reinforced. A high-fidelity Environment is the source of all meaningful experience, providing the rich, realistic feedback necessary for learning. This feedback is the lifeblood for the agent's Self-Evolution mechanism; without authentic data from a challenging world, adaptation stagnates, and the agent's internal model of reality becomes a fragile caricature. Finally, the quality of this entire feedback loop is mediated by the Coordination architecture. A sophisticated interface ensures that the agent's intentions are translated into effective actions and that environmental feedback is perceived without distortion, closing the loop and enabling the next cycle of evolution.

The interdependence is absolute: a brilliant self-evolution algorithm is useless if it learns from a simplistic world. A hyper-realistic world provides no benefit to an agent that cannot evolve. And a failure in coordination renders both a realistic world and an adaptive mind impotent.

The grand challenge for the next decade, therefore, is not simply to advance the state-of-the-art within each pillar, but to focus on the interfaces and feedback pathways that connect them. The most profound breakthroughs will come from research that asks: How can physical feedback from the body

reshape the brain's high-level plans? How can an agent's evolving self-awareness be used to actively seek out challenging scenarios in its environment? How can human interaction provide the richest form of feedback to accelerate the entire evolutionary loop?

We call for a community-wide focus to shift from optimizing isolated performance metrics to building and perfecting this complete, dynamic system. The most trustworthy embodied AI will not be the one with the single best component, but the one that achieves the most stable and adaptive harmony within this entire closed-loop system.

8 Conclusion

The field of Embodied Artificial Intelligence (EAI) is advancing at an unprecedented pace, with agents moving from simulated environments to complex, physical interactions in the real world. This increasing autonomy and physical capability, however, introduces profound challenges to safety and trustworthiness, where failures can result in direct physical harm, property damage, or the violation of societal norms. In this paper, we have provided a comprehensive framework to navigate this critical landscape. We began by establishing ten core principles, systematically organized under the two indispensable dimensions of Trustworthiness (accuracy, reliability, explainability, controllability, auditability) and Safety (attack resistance, abuse prevention, identifiability, privacy protection, value alignment). To unify disparate research efforts, we introduced a novel, agent-centric framework that analyzes risks across the four operational stages of an EAI system: Instruction Understanding, Environment Perception, Behavior Planning, and Physical Interaction. Within this structure, we systematically reviewed the current state-of-the-art, examining key solutions, benchmarks, evaluation metrics, and simulators, thereby identifying critical gaps and challenges. This paper has sought to move beyond a narrow focus on individual components, instead advocating for a holistic understanding of the entire EAI system. We conclude by reiterating our position that the future of safe and trustworthy EAI lies not in perfecting isolated modules, but in engineering the closed-loop, cybernetic system as a whole. Future progress hinges on a paradigm shift towards understanding the dynamic interplay between the Agent (Self), its Environment (World), and their Interaction. By focusing on this unified system, we can pave the way for the next generation of embodied agents that are not only capable and intelligent but are fundamentally safe and genuinely trustworthy.

References

[1] Mohammed Abugurain and Shinkyu Park. Integrating disambiguation and user preferences into large language models for robot motion planning. *arXiv preprint arXiv:2404.14547*, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Ananye Agarwal, Ashish Kumar, Jitendra Malik, and Deepak Pathak. Legged locomotion in challenging terrains using egocentric vision. In *Conference on robot learning*, pages 403–415, 2023.

[4] Matthias Althoff, Goran Frehse, and Antoine Girard. Set propagation techniques for reachability analysis. *Annual Review of Control, Robotics, and Autonomous Systems*, pages 369–395, 2021.

[5] Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4):15–15, 2007.

[6] Pasquale Antonante, David I Spivak, and Luca Carlone. Monitoring and diagnosability of perception systems. In *2021 IEEE/RSJ international conference on intelligent robots and systems*, pages 168–175, 2021.

[7] Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. *arXiv preprint arXiv:2409.13682*, 2024.

[8] Mattias Appelgren and Alex Lascarides. Interactive task learning via embodied corrective feedback. *Autonomous Agents and Multi-Agent Systems*, page 54, 2020.

[9] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, pages 136037–136083, 2024.

[10] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[11] IEEE Standards Association et al. Ieee standard for transparency of autonomous systems. *IEEE Std*, pages 7001–2021, 2022.

[12] Rumaisa Azeem, Andrew Hundt, Masoumeh Mansouri, and Martim Brandão. Llm-driven robots risk enacting discrimination, violence, and unlawful actions, 2024.

[13] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[14] Leonard Bärmann, Chad DeChant, Joana Plewnia, Fabian Peller-Konrad, Daniel Bauer, Tamim Asfour, and Alex Waibel. Episodic memory verbalization using hierarchical representations of life-long robot experience. *arXiv preprint arXiv:2409.17702*, 2024.

[15] Steve Benford, Eike Schneiders, Juan Pablo Martinez Avila, Praminda Caleb-Solly, Patrick Robert Brundell, Simon Castle-Green, Feng Zhou, Rachael Garrett, Kristina Höök, Sarah Whatley, et al. Somatic safety: An embodied approach towards safe human-robot interaction. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 429–438. IEEE, 2025.

[16] Kasra Borazjani, Payam Abdisarabshali, Fardis Nadimi, Naji Khosravan, Minghui Liwang, Xianbin Wang, Yiguang Hong, and Seyyedali Hosseinalipour. Multi-modal multi-task (m3t) federated foundation models for embodied ai: Potentials and challenges for edge integration. *arXiv preprint arXiv:2505.11191*, 2025.

[17] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. In *Artificial intelligence safety and security*, pages 57–69. Chapman and Hall/CRC, 2018.

[18] Yuliang Cai, Jesse Thomason, and Mohammad Rostami. Task-attentive transformer architecture for continual learning of vision-and-language tasks using knowledge distillation. *arXiv preprint arXiv:2303.14423*, 2023.

[19] Michael Cannon. An enactive approach to value alignment in artificial intelligence: A matter of relevance. In *Conference on Philosophy and Theory of Artificial Intelligence*, pages 119–135, 2021.

[20] Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. On the sensory commutativity of action sequences for embodied agents. *arXiv preprint arXiv:2002.05630*, 2020.

[21] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.

[22] Nicolas Harvey Chapman, Feras Dayoub, Will Browne, and Chris Lehnert. Enhancing embodied object detection through language-image pre-training and implicit object memory. *arXiv preprint arXiv:2402.03721*, 2024.

[23] Prithvijit Chattopadhyay, Judy Hoffman, Roozbeh Mottaghi, and Aniruddha Kembhavi. Robustnav: Towards benchmarking robustness in embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15700, 2021.

[24] Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. xrag: Extreme context compression for retrieval-augmented generation with one token. *Advances in Neural Information Processing Systems*, pages 109487–109516, 2024.

[25] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[26] Minkyu Choi, Yunhao Yang, Neel P Bhatt, Kushagra Gupta, Sahil Shah, Aditya Rai, David Fridovich-Keil, Ufuk Topcu, and Sandeep P Chinchali. Real-time privacy preservation for robot visual perception. *arXiv preprint arXiv:2505.05519*, 2025.

[27] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 2017.

[28] Paolo Cudrano, Xiaoyu Luo, and Matteo Matteucci. The empirical impact of forgetting and transfer in continual visual odometry, 2024.

[29] Min Dai and Aaron D Ames. Robust push recovery on bipedal robots: Leveraging multi-domain hybrid systems with reduced-order model predictive control. *arXiv preprint arXiv:2504.18698*, 2025.

[30] John M Davis, Ian A Gravagne, Billy J Jackson, and Robert J Marks II. Controllability, observability, realizability, and stability of dynamic linear systems. *arXiv preprint arXiv:0901.3764*, 2009.

[31] Francesco De Micco, Vittoradolfo Tambone, Paola Frati, Mariano Cingolani, and Roberto Scendoni. Disability 4.0: bioethical considerations on the use of embodied artificial intelligence. *Frontiers in Medicine*, 11:1437280, 2024.

[32] Francesco De Micco, Vittoradolfo Tambone, Paola Frati, Mariano Cingolani, and Roberto Scendoni. Disability 4.0: bioethical considerations on the use of embodied artificial intelligence. *Frontiers in Medicine*, 11:1437280, 2024.

[33] Fethiye Irmak Dogan, Maithili Patel, Weiyu Liu, Iolanda Leite, and Sonia Chernova. A model-agnostic approach for semantically driven disambiguation in human-robot interaction. *arXiv preprint arXiv:2409.17004*, 2024.

[34] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.

[35] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023.

[36] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[37] Kshitij Dwivedi, Gemma Roig, Aniruddha Kembhavi, and Roozbeh Mottaghi. What do navigation agents learn about their environment? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10276–10285, 2022.

[38] Amanuel Ergogo and Zhao Han. Towards embodied agent intent explanation in human-robot collaboration: Act error analysis and solution conceptualization. In *International Conference on Robotics and Automation 2025 Workshop: Human-Centered Robot Learning in the Era of Big Data and Large Models*, 2025.

[39] COMISSÃO EUROPEIA. Proposal for a regulation on a european approach for artificial intelligence. *Bruxelas: Comissão Europeia*, 2021.

[40] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[41] Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. *arXiv preprint arXiv:2501.00358*, 2024.

[42] Hongjie Fang, Hao-Shu Fang, Sheng Xu, and Cewu Lu. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *IEEE Robotics and Automation Letters*, 7(3):7383–7390, 2022.

[43] Xiaolin Fang, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Embodied uncertainty-aware object segmentation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2639–2646, 2024.

[44] Z Fang, A Jain, G Sarch, AW Harley, and K Fragkiadaki. Move to see better: Self-improving embodied object detection. arxiv 2020. *arXiv preprint arXiv:2012.00057*, 2020.

[45] MH Farhadi, Ali Rabiee, Sima Ghafoori, Anna Cetera, Wei Xu, and Reza Abiri. Human-centered shared autonomy for motor planning, learning, and control applications, 2025.

[46] Figure AI Inc. Helix: A vision-language-action model for generalist humanoid control. Technical report / blog post, February 2025.

[47] Masashi Fukunaga and Takeshi Sugawara. Random spoofing attack against lidar-based scan matching slam. *VehicleSec2024*, 2024.

[48] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

[49] Hang Gao and Yongfeng Zhang. Memory sharing for large language model based agents. *arXiv preprint arXiv:2404.09982*, 2024.

[50] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, pages 10049–10056, 2022.

[51] Xiaofeng Gao, Ran Gong, Tianmin Shu, Xu Xie, Shu Wang, and Song-Chun Zhu. Vrkitchen: an interactive 3d virtual environment for task-oriented learning. *arXiv preprint arXiv:1903.05757*, 2019.

[52] Artur d'Avila Garcez and Luis C Lamb. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, pages 12387–12406, 2023.

[53] Javier Garcia and Fernando Fernández. Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 45:515–564, 2012.

[54] Haoran Geng, Ziming Li, Yiran Geng, Jiayi Chen, Hao Dong, and He Wang. Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2978–2988, 2023.

[55] Liang Geng and Jianqin Yin. Viewinfer3d: 3d visual grounding based on embodied viewpoint inference. *IEEE Robotics and Automation Letters*, 2024.

[56] Reinhard Grabler and Sabine Theresia Koeszegi. Privacy beyond data: Assessment and mitigation of privacy risks in robotic technology for elderly care. *ACM Transactions on Human-Robot Interaction*, pages 1–23, 2025.

[57] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation*, pages 5021–5028, 2024.

[58] Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319:103905, 2023.

[59] Ji Guo, Long Zhou, Zhijin Wang, Jiaming He, Qiyang Song, Aiguo Chen, and Wenbo Jiang. Baddepth: Backdoor attacks against monocular depth estimation in the physical world, 2025.

[60] Dylan Hadfield-Menell, Anca D Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *The Association for the Advancement of Artificial Intelligence Workshops*, 2017.

[61] Hongmei He, John Gray, Angelo Cangelosi, Qinggang Meng, T Martin McGinnity, and Jörn Mehnen. The challenges and opportunities of human-centered ai for trustworthy robots and autonomous systems. *IEEE Transactions on Cognitive and Developmental Systems*, 14(4):1398–1412, 2021.

[62] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[63] Yaxin Hu, Arissa J Sato, Jingxin Du, Chenming Ye, Anjun Zhu, Pragathi Praveena, and Bilge Mutlu. Narraguide: an llm-based narrative mobile robot for remote place exploration. *arXiv preprint arXiv:2508.01235*, 2025.

[64] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023.

[65] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[66] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175, 2024.

[67] Xuying Huang, Sicong Pan, and Maren Bennewitz. Privacy risks of robot vision: A user study on image modalities and resolution. *arXiv preprint arXiv:2505.07766*, 2025.

[68] Xuying Huang, Sicong Pan, Olga Zatsarynna, Juergen Gall, and Maren Bennewitz. Improved semantic segmentation from ultra-low-resolution rgb images applied to privacy-preserving object-goal navigation. *arXiv preprint arXiv:2507.16034*, 2025.

[69] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[70] Yuting Huang, Leilei Ding, Zhipeng Tang, Tianfu Wang, Xinrui Lin, Wuyang Zhang, Mingxiao Ma, and Yanyong Zhang. A framework for benchmarking and aligning task-planning safety in llm-based embodied agents, 2025.

[71] Jeffrey Ichnowski, Yahav Avigal, Justin Kerr, and Ken Goldberg. Dex-nerf: Using a neural radiance field to grasp transparent objects. *arXiv preprint arXiv:2110.14217*, 2021.

[72] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[73] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, pages 3019–3026, 2020.

[74] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

[75] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, page 6, 2022.

[76] Eliot Krzysztof Jones, Alexander Robey, Andy Zou, Zachary Ravichandran, George J Pappas, Hamed Hassani, Matt Fredrikson, and J Zico Kolter. Adversarial attacks on robotic vision language action models. *arXiv preprint arXiv:2506.03350*, 2025.

[77] Rudolf E Kalman. On the general theory of control systems. In *Proceedings first international conference on automatic control, Moscow, USSR*, pages 481–492, 1960.

[78] Isaac Kasahara, Shubham Agrawal, Selim Engin, Nikhil Chavan-Dafle, Shuran Song, and Volkan Isler. Ric: Rotate-inpaint-complete for generalizable scene reconstruction. In *2024 IEEE International Conference on Robotics and Automation*, pages 2713–2720. IEEE, 2024.

[79] Osman Semih Kayhan, Bart Vredebregt, and Jan C. van Gemert. Hallucination in object detection – a study in visual part verification. *arXiv preprint arXiv:2106.02523*, 2021.

[80] Azal Ahmad Khan, Michael Andrev, Muhammad Ali Murtaza, Sergio Aguilera, Rui Zhang, Jie Ding, Seth Hutchinson, and Ali Anwar. Safety aware task planning via large language models in robotics. *arXiv preprint arXiv:2503.15707*, 2025.

[81] Junae Kim and Amardeep Kaur. A survey on adversarial robustness of lidar-based machine learning perception in autonomous vehicles. *arXiv preprint arXiv:2411.13778*, 2024.

[82] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[83] Myeung Un Kim, Harim Lee, Hyun Jong Yang, and Michael S Ryoo. Privacy-preserving robot vision with anonymized faces by extreme low resolution. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 462–467, 2019.

[84] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ international conference on intelligent robots and systems*, pages 2149–2154, 2004.

[85] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[86] Philip Koopman and Michael Wagner. Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*, pages 15–24, 2016.

[87] Klemen Kotar and Roozbeh Mottaghi. Interactron: Embodied adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14860–14869, 2022.

[88] Andrey Kurenkov, Michael Lingelbach, Tanmay Agarwal, Emily Jin, Chengshu Li, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, and Roberto Martın-Martın. Modeling dynamic environments with scene graph memory. In *International Conference on Machine Learning*, pages 17976–17993, 2023.

[89] In Lee. Service robots: a systematic literature review. *Electronics*, 10(21):2658, 2021.

[90] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[91] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, et al. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*, 2021.

[92] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, pages 100428–100534, 2024.

[93] Miao Li, Wenhao Ding, and Ding Zhao. Privacy risks in reinforcement learning for household robots. In *2024 IEEE International Conference on Robotics and Automation*, pages 5148–5154, 2024.

[94] Peng Li, Yupei Huang, Wenkai Chang, Chao Zhou, Shuo Wang, Junzhi Yu, and Zhengxing Wu. Active slam with dynamic viewpoint optimization for robust visual navigation. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[95] Rui Li, Zixuan Hu, Wenxi Qu, Jinouwen Zhang, Zhenfei Yin, Sha Zhang, Xuantuo Huang, Hanqing Wang, Tai Wang, Jiangmiao Pang, et al. Labutopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents. *arXiv preprint arXiv:2505.22634*, 2025.

[96] Yifan Li, Yuhang Chen, Anh Dao, Lichi Li, Zhongyi Cai, Zhen Tan, Tianlong Chen, and Yu Kong. Industryeqa: Pushing the frontiers of embodied question answering in industrial scenarios, 2025.

[97] Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, et al. Memos: An operating system for memory-augmented generation (mag) in large language models. *arXiv preprint arXiv:2505.22101*, 2025.

[98] Wenqi Liang, Gan Sun, Qian He, Yu Ren, Jiahua Dong, and Yang Cong. Never-ending behavior-cloning agent for robotic manipulation. *arXiv preprint arXiv:2403.00336*, 2024.

[99] Zhixuan Liang, Yao Mu, Hengbo Ma, Masayoshi Tomizuka, Mingyu Ding, and Ping Luo. Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16467–16476, 2024.

[100] Zihan Liao, Jun Wang, Hang Yu, Lingxiao Wei, Jianguo Li, and Wei Zhang. E2llm: Encoder elongated large language models for long-context understanding and reasoning. *arXiv preprint arXiv:2409.06679*, 2024.

[101] Aishan Liu, Zonghao Ying, Le Wang, Junjie Mu, Jinyang Guo, Jiakai Wang, Yuqing Ma, Siyuan Liang, Mingchuan Zhang, Xianglong Liu, et al. Agentsafe: Benchmarking the safety of embodied agents on hazardous instructions. *arXiv preprint arXiv:2506.14697*, 2025.

[102] Jinyi Liu, Yi Ma, Jianye Hao, Yujing Hu, Yan Zheng, Tangjie Lv, and Changjie Fan. Prioritized trajectory replay: A replay memory for data-driven reinforcement learning. *arXiv preprint arXiv:2306.15503*, 2023.

[103] Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. *arXiv preprint arXiv:2411.04999*, 2024.

[104] Shuyuan Liu, Jiawei Chen, Shouwei Ruan, Hang Su, and Zhaoxia Yin. Exploring the robustness of decision-level through adversarial attacks on llm-based embodied models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8120–8128, 2024.

[105] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403, 2024.

[106] Xinzhu Liu, Xinghang Li, Di Guo, Sinan Tan, Huaping Liu, and Fuchun Sun. Embodied multi-agent task planning from ambiguous instruction. In *Robotics: Science and Systems*, 2022.

[107] Yihao Liu, Xu Cao, Tingting Chen, Yankai Jiang, Junjie You, Minghua Wu, Xiaosong Wang, Mengling Feng, Yaochu Jin, and Jintai Chen. From screens to scenes: A survey of embodied ai in healthcare. *Information Fusion*, 119:103033, 2025.

[108] Xiaoxiao Long, Qingrui Zhao, Kaiwen Zhang, Zihao Zhang, Dingrui Wang, Yumeng Liu, Zhengjie Shu, Yi Lu, Shouzheng Wang, Xinzhe Wei, et al. A survey: Learning embodied intelligence from physical simulators and world models. *arXiv preprint arXiv:2507.00917*, 2025.

[109] Guanxing Lu, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. Thinkbot: Embodied instruction following with thought chain reasoning. *arXiv preprint arXiv:2312.07062*, 2023.

[110] Liming Lu, Shuchao Pang, Siyuan Liang, Haotian Zhu, Xiyu Zeng, Aishan Liu, Yunhuai Liu, and Yongbin Zhou. Adversarial training for multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2503.04833*, 2025.

[111] Xiaoya Lu, Zeren Chen, Xuhao Hu, Yijin Zhou, Weichen Zhang, Dongrui Liu, Lu Sheng, and Jing Shao. Is-bench: Evaluating interactive safety of vlm-driven embodied agents in daily household tasks. *arXiv preprint arXiv:2506.16402*, 2025.

[112] Xuancun Lu, Zhengxian Huang, Xinfeng Li, Wenyuan Xu, et al. Poex: Understanding and mitigating policy executable jailbreak attacks against embodied ai. *arXiv preprint arXiv:2412.16633*, 2024.

[113] Ewa Luger and Abigail Sellen. " like having a really bad pa" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 conference on human factors in computing systems*, pages 5286–5297, 2016.

[114] Wenqi Lyu, Zerui Li, Yanyuan Qiao, and Qi Wu. Badnaver: Exploring jailbreak attacks on vision-and-language navigation. *arXiv preprint arXiv:2505.12443*, 2025.

[115] Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18081–18090, 2024.

[116] Connor Malone, Owen Claxton, Iman Shames, and Michael Milford. Adversarial attacks and detection in visual place recognition for safer robot navigation, 2025.

[117] Junyuan Mao, Fanci Meng, Yifan Duan, Miao Yu, Xiaojun Jia, Junfeng Fang, Yuxuan Liang, Kun Wang, and Qingsong Wen. Agentsafe: Safeguarding large language model-based multi-agent systems via hierarchical data management, 2025.

[118] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, pages 7327–7334, 2022.

[119] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys*, 54(6):1–35, 2021.

[120] Jorge Mendez-Mendez, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Embodied lifelong learning for task and motion planning. In *Conference on Robot Learning*, pages 2134–2150, 2023.

[121] Catherine Menon, Austen Rainer, Patrick Holthaus, Gabriella Lakatos, and Silvio Carta. Ehazop: A proof of concept ethical hazard analysis of an assistive robot. *arXiv preprint arXiv:2406.09239*, 2024.

[122] Leila Methnani, Manolis Chiou, Virginia Dignum, and Andreas Theodorou. Who's in charge here? a survey on trustworthy ai in variable autonomy robotic systems. *ACM computing surveys*, 56(7):1–32, 2024.

[123] Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science robotics*, page eabk2822, 2022.

[124] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, pages 1–38, 2019.

[125] So Yeon Min, Xavi Puig, Devendra Singh Chaplot, Tsung-Yen Yang, Akshara Rai, Priyam Parashar, Ruslan Salakhutdinov, Yonatan Bisk, and Roozbeh Mottaghi. Situated instruction following. In *European Conference on Computer Vision*, pages 202–228, 2024.

[126] John Molloy and John McDermid. Safety assessment for autonomous systems' perception capabilities. *arXiv preprint arXiv:2208.08237*, 2022.

[127] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, pages 25081–25094, 2023.

[128] Adithyavairavan Murali, Arsalan Mousavian, Clemens Eppner, Chris Paxton, and Dieter Fox. 6-dof grasping for target-driven object manipulation in clutter. In *2020 IEEE International Conference on Robotics and Automation*, pages 6232–6238. IEEE, 2020.

[129] Robin R. Murphy. Introduction to ai robotics. *Robotica*, pages 569–571, 2002.

[130] Rokuto Nagata, Kenji Koide, Yuki Hayakawa, Ryo Suzuki, Kazuma Ikeda, Ozora Sako, Qi Alfred Chen, Takami Sato, and Kentaro Yoshioka. Slamspoof: Practical lidar spoofing attacks on localization systems guided by scan matching vulnerability analysis. *arXiv preprint arXiv:2502.13641*, 2025.

[131] Mohaiminul Al Nahian, Zainab Altaweel, David Reitano, Sabbir Ahmed, Shiqi Zhang, and Adnan Siraj Rakin. Robo-troj: Attacking llm-based task planners. *arXiv preprint arXiv:2504.17070*, 2025.

[132] I Nahrendra, Byeongho Yu, Minho Oh, Dongkyu Lee, Seunghyun Lee, Hyeonwoo Lee, Hyungtae Lim, and Hyun Myung. Obstacle-aware quadrupedal locomotion with resilient multi-modal reinforcement learning. *arXiv preprint arXiv:2409.19709*, 2024.

[133] Nithesh Naik, BM Hameed, Dasharathraj K Shetty, Dishant Swain, Milap Shah, Rahul Paul, Kaivalya Aggarwal, Sufyan Ibrahim, Vathsala Patil, Komal Smriti, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Frontiers in surgery*, 9:862322, 2022.

[134] Subash Neupane, Shaswata Mitra, Ivan A Fernandez, Swayamjit Saha, Sudip Mittal, Jingdao Chen, Nisha Pillai, and Shahram Rahimi. Security considerations in ai-robotics: A survey of current methods, challenges, and opportunities. *IEEE Access*, 12:22072–22097, 2024.

[135] Celia Nieto Agraz, Pascal Hinrichs, Marco Eichelberg, and Andreas Hein. Is the robot spying on me? a study on perceived privacy in telepresence scenarios in a care setting with mobile and humanoid robots. *International Journal of Social Robotics*, pages 1–15, 2024.

[136] David Nilsson, Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Embodied visual active learning for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2373–2383, 2021.

[137] Nvidia. Nvidia isaac sim: Robotics simulation and synthetic data.

[138] Joanna Isabelle Olszewska, Data Compression Standard Committee, et al. Ieee standard for data privacy process. *IEEE Std 7002-2022*, pages 1–41, 2022.

[139] Laurent Orseau and M Armstrong. Safely interruptible agents. In *Conference on Uncertainty in Artificial Intelligence*. Association for Uncertainty in Artificial Intelligence, 2016.

[140] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, pages 27730–27744, 2022.

[141] Cheng Pan, Kai Junge, and Josie Hughes. Vision-language-action model and diffusion policy switching enables dexterous control of an anthropomorphic hand. *arXiv preprint arXiv:2410.14022*, 2024.

[142] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[143] Jonghyuk Park, Alex Lascarides, and Subramanian Ramamoorthy. Learning visually grounded domain ontologies via embodied conversation and explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14361–14368, 2025.

[144] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[145] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022.

[146] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics*, pages 13387–13434, 2023.

[147] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation*, pages 3406–3413. IEEE, 2016.

[148] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1988.

[149] Pradip Pramanick, Chayan Sarkar, Sayan Paul, Ruddra dev Roychoudhury, and Brojeshwar Bhowmick. Doro: Disambiguation of referred object for embodied agents. *IEEE Robotics and Automation Letters*, pages 10826–10833, 2022.

[150] Sai Prasanna, Daniel Honerkamp, Kshitij Sirohi, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. Perception matters: Enhancing embodied ai with uncertainty-aware semantic segmentation. *arXiv preprint arXiv:2408.02297*, 2024.

[151] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018.

[152] Yong Qi, Gabriel Kyebambo, Siyuan Xie, Wei Shen, Shenghui Wang, Bitao Xie, Bin He, Zhipeng Wang, and Shuo Jiang. Safety control of service robots with llms and embodied knowledge graphs. *arXiv preprint arXiv:2405.17846*, 2024.

[153] Benedict Quartey, Eric Rosen, Stefanie Tellex, and George Konidaris. Verifiably following complex robot instructions with foundation models. *arXiv preprint arXiv:2402.11498*, 2024.

[154] Frano Rajič. Robustness of embodied point navigation agents. In *European Conference on Computer Vision*, pages 193–204, 2022.

[155] Ram Ramrakhya, Matthew Chang, Xavier Puig, Ruta Desai, Zsolt Kira, and Roozbeh Mottaghi. Grounding multimodal llms to embodied agents that ask for help with reinforcement learning. *arXiv preprint arXiv:2504.00907*, 2025.

[156] Ram Ramrakhya, Matthew Chang, Xavier Puig, Ruta Desai, Zsolt Kira, and Roozbeh Mottaghi. Grounding multimodal llms to embodied agents that ask for help with reinforcement learning, 2025.

[157] Ram Ramrakhya, Eric Undersander, Dhruv Batra, and Abhishek Das. Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5173–5183, 2022.

[158] Siddharth Reddy, Anca D Dragan, and Sergey Levine. Shared autonomy via deep reinforcement learning, 2018.

[159] Protection Regulation. General data protection regulation. *Intouch*, pages 1–5, 2018.

[160] Adrian Remonda, Cole Corbitt Terrell, Eduardo E Veas, and Marc Masana. Uncertainty-based experience replay for task-agnostic continual reinforcement learning. *Transactions on Machine Learning Research*, 2025.

[161] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J Pappas. Jailbreaking llm-controlled robots. *arXiv preprint arXiv:2410.13691*, 2024.

[162] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635, 2011.

[163] Gabriel Sarch, Yue Wu, Michael J Tarr, and Katerina Fragkiadaki. Open-ended instructable embodied agents with memory-augmented large language models. *arXiv preprint arXiv:2310.15127*, 2023.

[164] Chayan Sarkar, Avik Mitra, Pradip Pramanick, and Tapas Nayak. tage: Enabling an embodied agent to understand human instructions. *arXiv preprint arXiv:2310.15605*, 2023.

[165] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. *arXiv preprint arXiv:1803.00653*, 2018.

[166] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019.

[167] Stefan Schaal. Learning from demonstration. *Advances in neural information processing systems*, 1996.

[168] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, pages 233–242, 1999.

[169] Mingyo Seo, Ryan Gupta, Yifeng Zhu, Alexy Skoutnev, Luis Sentis, and Yuke Zhu. Learning to walk by steering: Perceptive quadrupedal locomotion in dynamic environments. In *2023 IEEE International Conference on Robotics and Automation*, 2023.

[170] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics: Results of the 11th international conference*, pages 621–635. Springer, 2017.

[171] Lingdong Shen, Chunlei Huo, Nuo Xu, Chaowei Han, and Zichen Wang. Learn how to see: collaborative embodied learning for object detection and camera adjusting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4793–4801, 2024.

[172] Yide Shentu, Philipp Wu, Aravind Rajeswaran, and Pieter Abbeel. From llms to actions: Latent codes as bridges in hierarchical robot control. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 8539–8546. IEEE, 2024.

[173] Sangwoo Shin, Seunghyun Kim, Youngsoo Jang, Moontae Lee, and Honguk Woo. Semantic skill grounding for embodied instruction-following in cross-domain environments. *arXiv preprint arXiv:2408.01024*, 2024.

[174] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799, 2023.

[175] Kunal Pratap Singh, Luca Weihs, Alvaro Herrasti, Jonghyun Choi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Ask4help: Learning to leverage an expert for embodied tasks. *Advances in Neural Information Processing Systems*, pages 16221–16232, 2022.

[176] Ida Skubis, Agata Mesjasz-Lech, and Joanna Nowakowska-Grunt. Humanoid robots in tourism and hospitality—exploring managerial, ethical, and societal challenges. *Applied Sciences*, 14(24):11823, 2024.

[177] Yejin Son, Minseo Kim, Sungwoong Kim, Seungju Han, Jian Kim, Dongju Jang, Youngjae Yu, and Chanyoung Park. Subtle risks, critical failures: A framework for diagnosing physical safety of llms for embodied decision making, 2025.

[178] Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters*, 2024.

[179] Pablo Sprechmann, Siddhant M Jayakumar, Jack W Rae, Alexander Pritzel, Adria Puigdomenech Badia, Benigno Uria, Oriol Vinyals, Demis Hassabis, Razvan Pascanu, and Charles Blundell. Memory-based parameter adaptation. *arXiv preprint arXiv:1802.10542*, 2018.

[180] DiJia Su, Sainbayar Sukhbaatar, Michael Rabbat, Yuandong Tian, and Qinqing Zheng. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces, 2024.

[181] Shibo Sun, Xue Li, Donglin Di, Mingjie Wei, Lanshun Nie, Wei-Nan Zhang, Dechen Zhan, Yang Song, and Lei Fan. Llapa: A vision-language model framework for counterfactual-aware procedural planning. *arXiv preprint arXiv:2507.08496*, 2025.

[182] Yiliu Sun, Yanfang Zhang, Zicheng Zhao, Sheng Wan, Dacheng Tao, and Chen Gong. Fast-slow-thinking: Complex task solving with large language models, 2025.

[183] Yitong Sun, Yao Huang, and Xingxing Wei. Embodied laser attack: leveraging scene priors to achieve agent-based robust non-contact attacks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5902–5910, 2024.

[184] Youbang Sun, Xiang Wang, Jie Fu, Chaochao Lu, and Bowen Zhou. $\mathtt{R^2AI}$: Towards resistant and resilient ai in an evolving world. *arXiv preprint arXiv:2509.06786*, 2025.

[185] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation*, pages 13438–13444. IEEE, 2021.

[186] Mohammed Tahri Sqalli, Begali Aslonov, Mukhammadjon Gafurov, and Shokhrukhbek Nurmatov. Humanizing ai in medical training: ethical framework for responsible design. *Frontiers in Artificial Intelligence*, 6:1189914, 2023.

[187] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[188] Russ Tedrake, Ian R Manchester, Mark Tobenkin, and John W Roberts. Lqr-trees: Feedback motion planning via sums-of-squares verification. *The International Journal of Robotics Research*, pages 1038–1052, 2010.

[189] Jesse Thomason, Aishwarya Padmakumar, Jivko Sinapov, Nick Walker, Yuqian Jiang, Harel Yedidsion, Justin Hart, Peter Stone, and Raymond J Mooney. Improving grounded natural language understanding through human-robot dialog. In *2019 International Conference on Robotics and Automation*, pages 6934–6941, 2019.

[190] Ran Tian, Liting Sun, Andrea Bajcsy, Masayoshi Tomizuka, and Anca D Dragan. Safety assurances for human-robot interaction via confidence-aware game-theoretic human models. In *2022 International Conference on Robotics and Automation*, pages 11229–11235, 2022.

[191] Emanuel Todorov and Michael I Jordan. Optimal feedback control as a theory of motor coordination. *Nature neuroscience*, pages 1226–1235, 2002.

[192] Tristan Tomilin, Meng Fang, and Mykola Pechenizkiy. Hasard: A benchmark for vision-based safe reinforcement learning in embodied agents, 2025.

[193] Jim Torresen. A review of future and ethical perspectives of robotics and ai. *Frontiers in Robotics and AI*, 4:75, 2018.

[194] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. *arXiv preprint arXiv:2209.03855*, 2022.

[195] Víctor Mayoral Vilches, Laura Alzola Kirschgens, Asier Bilbao Calvo, Alejandro Hernández Cordero, Rodrigo Izquierdo Pisón, David Mayoral Vilches, Aday Muñiz Rosas, Gorka Olalde Mendia, Lander Usategi San Juan, Irati Zamalloa Ugarte, et al. Introducing the robot security framework (rsf), a standardized methodology to perform security assessments in robotics. *arXiv preprint arXiv:1806.04042*, 2018.

[196] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction*, 10(3):1–24, 2021.

[197] Laura Waltersdorfer, Fajar J Ekaputra, Tomasz Miksa, and Marta Sabou. Auditmai: Towards an infrastructure for continuous ai auditing. *arXiv preprint arXiv:2406.14243*, 2024.

[198] Hanlin Wang, Chak Tou Leong, Jian Wang, and Wenjie Li. E2cl: exploration-based error correction learning for embodied agents. *arXiv preprint arXiv:2409.03256*, 2024.

[199] Hanqing Wang, Jiahe Chen, Wensi Huang, Qingwei Ben, Tai Wang, Boyu Mi, Tao Huang, Siheng Zhao, Yilun Chen, Sizhe Yang, et al. Grutopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943*, 2024.

[200] Haoyu Wang, Christopher M Poskitt, and Jun Sun. Agentspec: Customizable runtime enforcement for safe and reliable llm agents. *arXiv preprint arXiv:2503.18666*, 2025.

[201] Ning Wang, Zihan Yan, Weiyang Li, Chuan Ma, He Chen, and Tao Xiang. Advancing embodied agent security: From safety benchmarks to input moderation. *arXiv preprint arXiv:2504.15699*, 2025.

[202] Pengyu Wang, Jialu Li, and Ling Shi. Optimal actuator attacks on autonomous vehicles using reinforcement learning. *arXiv preprint arXiv:2502.07839*, 2025.

[203] Sicheng Wang, Milutin N Nikolić, Tin Lun Lam, Qing Gao, Runwei Ding, and Tianwei Zhang. Robot manipulation based on embodied visual perception: A survey. *CAAI Transactions on Intelligence Technology*, 2025.

[204] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024.

[205] Xiaofei Wang, Mingliang Han, Tianyu Hao, Cegang Li, Yunbo Zhao, and Keke Tang. Advgrasp: Adversarial attacks on robotic grasping from a physical perspective. *arXiv preprint arXiv:2507.09857*, 2025.

[206] Xiaohan Wang, Yuehu Liu, Xinhang Song, Beibei Wang, and Shuqiang Jiang. Generating explanations for embodied action decision from visual observation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2838–2846, 2023.

[207] Yuanfei Wang, Xinju Huang, Fangwei Zhong, Yaodong Yang, Yizhou Wang, Yuanpei Chen, and Hao Dong. From strangers to assistants: Fast desire alignment for embodied agent-user adaptation. *arXiv preprint arXiv:2505.22503*, 2025.

[208] Zixuan Wang, Bo Yu, Junzhe Zhao, Wenhao Sun, Sai Hou, Shuai Liang, Xing Hu, Yinhe Han, and Yiming Gan. Karma: Augmenting embodied ai agents with long-and-short term memory systems. *arXiv preprint arXiv:2409.14908*, 2024.

[209] Maciej Wołczyk, Michał Zając, Razvan Pascanu, Łukasz Kuciński, and Piotr Miłoś. Continual world: A robotic benchmark for continual reinforcement learning. *Advances in Neural Information Processing Systems*, pages 28496–28510, 2021.

[210] Lik Hang Kenny Wong, Xueyang Kang, Kaixin Bai, and Jianwei Zhang. A survey of robotic navigation and manipulation with physics simulators in the era of embodied ai. *arXiv preprint arXiv:2505.01458*, 2025.

[211] Chengwei Wu, Weiran Yao, Wensheng Luo, Wei Pan, Guanghui Sun, Hui Xie, and Ligang Wu. A secure robot learning framework for cyber attack scheduling and countermeasure. *IEEE Transactions on Robotics*, 39(5):3722–3738, 2023.

[212] Di Wu, Jiaxin Fan, Junzhe Zang, Guanbo Wang, Wei Yin, Wenhao Li, and Bo Jin. Reinforced reasoning for embodied planning. *arXiv preprint arXiv:2505.22050*, 2025.

[213] Lingxuan Wu, Xiao Yang, Yinpeng Dong, Liuwei Xie, Hang Su, and Jun Zhu. Embodied active defense: Leveraging recurrent feedback to counter adversarial patches. *arXiv preprint arXiv:2404.00540*, 2024.

[214] Tao Wu, Chuhao Zhou, Yen Heng Wong, Lin Gu, and Jianfei Yang. Noisyeqa: Benchmarking embodied question answering against noisy queries. *arXiv preprint arXiv:2412.10726*, 2024.

[215] Wenxi Wu, Fabio Pierazzi, Yali Du, and Martim Brandão. Characterizing physical adversarial attacks on robot motion planners. In *2024 IEEE International Conference on Robotics and Automation*, pages 14319–14325, 2024.

[216] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Hang Yin, Yinan Liang, Angyuan Ma, Jiwen Lu, and Haibin Yan. Embodied instruction following in unknown environments. *arXiv preprint arXiv:2406.11818*, 2024.

[217] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.

[218] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[219] Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv preprint arXiv:2409.18313*, 2024.

[220] Quanting Xie, Shunyu So, Sachin Gupta, Zan Kotar, Berivan Eyuboglu, et al. Embodied-rag: General non-parametric embodied memory for retrieval and generation. *arXiv preprint arXiv:2409.18313*, 2024.

[221] Wenpeng Xing, Minghao Li, Mohan Li, and Meng Han. Towards robust and secure embodied ai: A survey on vulnerabilities and attacks. *arXiv preprint arXiv:2502.13175*, 2025.

[222] Zikang Xiong and Suresh Jagannathan. Manipulating neural path planners via slight perturbations. *IEEE Robotics and Automation Letters*, 9(6):5006–5013, 2024.

[223] Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles, 2022.

[224] Yuan Xu, Xingshuo Han, Gelei Deng, Jiwei Li, Yang Liu, and Tianwei Zhang. Sok: Rethinking sensor spoofing attacks against robotic vehicles from a systematic view. In *2023 IEEE 8th European Symposium on Security and Privacy*, pages 1082–1100, 2023.

[225] Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. Chalet: Cornell house agent learning environment. *arXiv preprint arXiv:1801.07357*, 2018.

[226] Cheng-Fu Yang, Yen-Chun Chen, Jianwei Yang, Xiyang Dai, Lu Yuan, Yu-Chiang Frank Wang, and Kai-Wei Chang. Lacma: Language-aligning contrastive learning with meta-actions for embodied instruction following. *arXiv preprint arXiv:2310.12344*, 2023.

[227] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[228] Jirui Yang, Zheyu Lin, Shuhan Yang, Zhihui Lu, and Xin Du. Concept enhancement engineering: A lightweight and efficient robust defense against jailbreak attacks in embodied ai. *arXiv preprint arXiv:2504.13201*, 2025.

[229] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. Snapmem: Snapshot-based 3d scene memory for embodied exploration and reasoning. *CoRR*, 2024.

[230] Ziyi Yang, Shreyas S Raman, Ankit Shah, and Stefanie Tellex. Plug in the safety chip: Enforcing constraints for llm-driven robot agents. In *2024 IEEE International Conference on Robotics and Automation*, pages 14435–14442, 2024.

[231] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[232] Mohammad Hasan Yeganegi, Majid Khadiv, S Ali A Moosavian, Jia-Jie Zhu, Andrea Del Prete, and Ludovic Righetti. Robust humanoid locomotion using trajectory optimization and sample-efficient learning. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots*, pages 170–177. IEEE, 2019.

[233] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, et al. Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565*, 2023.

[234] Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. Safeagentbench: A benchmark for safe task planning of embodied llm agents. *arXiv preprint arXiv:2412.13178*, 2024.

[235] Ziyi Yin, Yuanpu Cao, Han Liu, Ting Wang, Jinghui Chen, and Fenhlong Ma. Towards robust multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2502.00653*, 2025.

[236] Pian Yu, Shuyang Dong, Shili Sheng, Lu Feng, and Marta Kwiatkowska. Trust-aware motion planning for human-robot collaboration under distribution temporal logic specifications. In *2024 IEEE International Conference on Robotics and Automation*, pages 12949–12955, 2024.

[237] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100, 2020.

[238] Tianjiao Yu, Vedant Shah, Muntasir Wahed, Kiet A Nguyen, Adheesh Juvekar, Tal August, and Ismini Lourentzou. Uncertainty in action: Confidence elicitation in embodied agents. *arXiv preprint arXiv:2503.10628*, 2025.

[239] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Jiahao Xu, Tian Liang, Pinjia He, and Zhaopeng Tu. Refuse whenever you feel unsafe: Improving safety in llms via decoupled refusal training. *arXiv preprint arXiv:2407.09121*, 2024.

[240] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

[241] Kevin Zakka, Andy Zeng, Johnny Lee, and Shuran Song. Form2fit: Learning shape priors for generalizable assembly from disassembly. In *2020 IEEE International Conference on Robotics and Automation*, pages 9404–9410. IEEE, 2020.

[242] Chong Zhang, Jin Jin, Jonas Frey, Nikita Rudin, Matías Mattamala, Cesar Cadena, and Marco Hutter. Resilient legged local navigation: Learning to traverse with compromised perception end-to-end. In *2024 IEEE International Conference on Robotics and Automation*, pages 34–41, 2024.

[243] Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, et al. Badrobot: Jailbreaking embodied llms in the physical world. *arXiv preprint arXiv:2407.20242*, 2024.

[244] Hongxin Zhang, Zheyuan Zhang, Zeyuan Wang, Zunzhe Zhang, Lixing Fang, Qinhong Zhou, and Chuang Gan. Ella: Embodied social agents with lifelong memory. *arXiv preprint arXiv:2506.24019*, 2025.

[245] Jenny Zhang, Samson Yu, Jiafei Duan, and Cheston Tan. Good time to ask: A learning framework for asking for help in embodied visual navigation. In *2023 20th International Conference on Ubiquitous Robots*, pages 503–509, 2023.

[246] Zhipeng Zhang, Zhimin Wei, Guolei Sun, Peng Wang, and Luc Van Gool. Self-explainable affordance learning with embodied caption. *arXiv preprint arXiv:2404.05603*, 2024.

[247] Han Zheng, Jiale Zhang, Mingyang Jiang, Peiyuan Liu, Danni Liu, Tong Qin, and Ming Yang. Embodied escaping: End-to-end reinforcement learning for robot navigation in narrow environment. *arXiv preprint arXiv:2503.03208*, 2025.

[248] Ying Zheng, Lei Yao, Yuejiao Su, Yi Zhang, Yi Wang, Sicheng Zhao, Yiyi Zhang, and Lap-Pui Chau. A survey of embodied learning for object-centric robotic manipulation. *Machine Intelligence Research*, pages 1–39, 2025.

[249] Fangwei Zhong, Kui Wu, Churan Wang, Hao Chen, Hai Ci, Zhoujun Li, and Yizhou Wang. Unrealzoo: Enriching photo-realistic virtual worlds for embodied ai. *arXiv preprint arXiv:2412.20977*, 2024.

[250] Xian Zhou, Yiling Qiao, Zhenjia Xu, TH Wang, Z Chen, J Zheng, Z Xiong, Y Wang, M Zhang, P Ma, et al. Genesis: A generative and universal physics engine for robotics and beyond. *arXiv preprint arXiv:2401.01454*, 2024.

[251] Xueyang Zhou, Guiyao Tie, Guowen Zhang, Hechang Wang, Pan Zhou, and Lichao Sun. Badvla: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization. *arXiv preprint arXiv:2505.16640*, 2025.

[252] Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*, 2025.

[253] Zihao Zhu, Bingzhe Wu, Zhengyou Zhang, Lei Han, Qingshan Liu, and Baoyuan Wu. Earbench: Towards evaluating physical risk awareness for task planning of foundation model-based embodied ai agents, 2024.

[254] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183, 2023.

[255] René Zurbrügg, Hermann Blum, Cesar Cadena, Roland Siegwart, and Lukas Schmid. Embodied active domain adaptation for semantic segmentation via informative path planning. *IEEE Robotics and Automation Letters*, pages 8691–8698, 2022.